(* TWO-WAY *)

18

Expression or Extension?
(Two Uses of Comparison Values)

Introduction 18-1

18A. East Coast climate again 18-3
clarity, at 18-11

18B. How much change of expression? 18-14
comparison values, at 18-16

18C. Seizing a better expression
- catch of plankton 18-24

18D. Feeling around for a better expression
- control of insects 18-30
- trying to squeeze closer 18-44

18E. Extension of the fit
- East Coast climate 18-47

18F. Pictures of expanded fits 18-55
- back to the East Coast 18-58

18G. Diagnostic plots need resistant analyses
- dry/wet ratios 18-65

****

18H. How far have we come? 18-76

****

18K. An interpretation (background) 18-78

****

Exh.		Page
1	A	18-7/9
2		18-10
3		18-13
4	B	18-15
5		18-18
6		18-19
7		18-20
8		18-21
9		18-22
10	C	18-25/26
11		18-28/29
12	D	18-31
13		18-32/33
14		18-34
15		18-36/37
16		18-38
17		18-39/41
18		18-42
19		18-43
20		18-45
21		18-46
22	E	18-48/49
23		18-50
24		18-52
25		18-53/54
26	F	18-56
27		18-59
28		18-61
29		18-62/63
30		18-64
31	G	18-66
32		18-68
33		18-69
34		18-71/72
35		18-73
36		18-75
	H	
	K	
	P	

18

Expression or Extension?

(Two Uses of Comparison Values)

One approach to any partly satisfactory analysis is to ask about changes in expression. We have seen this in making batches more symmetric, in making sets of batches more equally spread, and in making plots of relationship more nearly straight. We ought to expect it in making the residuals of a two-way table more nearly negligible.

Another approach is to complicate the structure of the fit somewhat. We did not see how to do this with batches, though we might come close by finding and using some way to describe amounts of unsymmetry. We saw examples of this when we began to straighten and flatten plots of relationship by fitting bend as well as tilt. We ought not to be surprised if we find that examples of this approach offer a way to go further with two-way tables.

Which approach when? This is a question that the data actually being analyzed may not be enough to answer. Consistency across quite different sets of data involving the same response may, for instance, be a controlling feature. To find, for example, that we need to use sometimes the raw value, sometimes the root, and sometimes the log of the same response could be less than satisfactory, especially

when nothing in the broad character and general structure of each particular body of data makes it clear which should be used for it. We may need a lot of data to become sure.

How we analyze data -- if we choose well -- has much to do with how we ought to think about it.

If there is a broadly useful way to express some response -- as by taking logs of human blood pressures or logs of stock prices -- we ought to use it, and make things simpler across a broad range. To use any other approach would be to lose simplicity and a possibility of deft thinking -- either would be a great loss (If a successful change of expression looks complex and peculiar when written as a formula, we can always forget the formula -- using a table or making up special graph paper.)

If no one choice of expression is broadly useful, although some very simple choice works in nearly every instance -- or even always -- we ought to be reluctant to adopt the approach of changing expression from instance to instance. Doing this would make it hard to compare results of analyses of different sets of data. And it would tend to make our thinking muzzy and confused. We dare not pay such prices in the long run.

We could take such caution too far. There are responses that need to be expressed in different ways for clear reasons. Sometimes in two ways within a single

problem. The situations that then arise are messy and hard to think through. Such difficulties, however, can be essential. Swapping back and forth between two expressions -- hard though it may be -- can be the easiest way to understand matters.

The optics of multi-element lenses -- those in good cameras, for instance -- is complicated for just this sort of reason. (Lengths need to be raw when we move from one element to the next, but reciprocal when we are concerned with what is focused where.) There are a variety of other examples. Such situations are still, fortunately, the exception rather than the rule.

Generally, we want -- once we have settled down -- to work with one expression of a single response. To do this we will often prefer to accept some complication of the structures we choose to fit.

In the beginning, though, when we attack a new kind of response, we ought to feel free to go as far as we can by choosing its expression. We ought also to be prepared to change gears -- from changed expression to expanded fit -- when, as, and only if this appears to be better.

18A. East Coast Climate Again

The most unsatisfactory elementary analysis we have so far looked at is our analysis of East Coast

monthly temperatures. (See exhibit 12 of chapter 16.) The residuals are both large and very orderly. We clearly ought to see whether a better choice of expression will help us.

The largest residuals were at or nearly at the four corners of the two-way plot. Accordingly, we might well begin by looking at just 4 monthly temperatures, one for each combination of

hottest month or coldest month

with

hottest place or coldest place

The four numbers are:

	January	July
Caribou	8.7	64.0
Laredo	57.6	87.7

These numbers make our failure to get a close description with a simple elementary analysis clear. From January to July is 55.3 degrees warmer in Caribou, but only 30.1 degrees warmer in Laredo. Or, if you like it the other way, Laredo is 48.9 degrees warmer than Caribou

in January, but only 23.7 degrees warmer in July. Whichever way we look at it, we fail by

$$55.3 - 30.1 = \mathbf{25.2} = 48.9 - 23.7$$

degrees to have the result of changing one factor the same at the extremes of the other.

What if we were to change expression? At present the change is smaller for the warmer version -- whether warmer month or warmer place. To move toward roots and logs would accentuate this difference -- would mean going the wrong way. What about squares instead? (We shall postpone the question of which of these new choices might indeed be sensible until later.)

If we take the squares of temperatures in degrees Fahrenheit, we find:

	January	July
Caribou	76	4096
Laredo	3318	7691

July minus January is now 4020 in Caribou and 4643 in Laredo. The Laredo change is the Caribou change plus 15%, where in raw temperature it was Caribou minus 45%. Squares seem to overcorrect a little, but seem to

come much closer to balance. So far as these four points go, squares seem worth a try.

Exhibit 1 shows the full table of squared temperatures, and its elementary analysis. To avoid unnecessary detail, we have chosen 10 $(°F)^2$ as the unit. The calculations are standard. Exhibit 2 shows the two-way plot of the fitted temperature. Notice that we have given a vertical scale in terms of temperature, not squared temperature. As a result, the vertical scale is far from uniform.

Why did we go back to temperature? Not because the raw data came to us in temperatures. We would certainly not always go back to the raw form. One or both of two reasons could send us back:

- a belief that people do think in °F.
- a belief that people ought to think in °F.

Both apply here. Either or neither may apply in other cases.

We have managed a better fit than we had in terms of raw temperature, as we can see from

	Laredo MINUS Caribou	July MINUS January	midspread
Full table in °F*	37.1	40.1	3.9
Exhibits 1 and 2	387	425	26.5

* Will appear later as exhibit 22.

The midspread of the residuals has been reduced from 10%

exhibit 1 of chapter 18

Squared monthly temperatures for seven East-coast airports

(raw data will come in exhibit 22)

A) The DATA -- in units of 10 $(°F)^2$

	Laredo	Baton Rouge	Atlanta	Wash'n	Boston	Portland	Caribou
Jan	332	275	199	131	85	43	8
Feb	383	309	218	138	85	46	10
Mar	468	364	278	205	141	102	47
Apr	576	446	381	296	223	176	120
May	659	542	490	419	334	274	235
June	736	634	604	539	452	382	341
July	769	658	632	598	521	460	410
Aug	771	651	618	569	511	441	382
Sept	697	601	554	484	413	343	283
Oct	585	484	402	339	302	234	177
Nov	436	349	269	228	197	141	85
Dec	352	287	204	144	108	63	21

B) ELEMENTARY ANALYSIS - same units

	Laredo	Baton Rouge	Atlanta	Wash'n	Boston	Portl	Caribou	Effect	Mean
Jan	-34	8	-7	-12	2	16	29	-198	153
Feb	0	25	-5	-22	-15	2	14	-181	170
Mar	26	21	-4	-14	-18	-1	-8	-122	229
Apr	46	15	11	-11	-24	-15	-23	-34	317
May	26	6	15	7	-18	-22	-13	71	422
June	-4	-7	24	22	-5	-19	-12	176	527
July	-22	-34	1	30	13	8	6	227	578
Aug	-5	-26	2	16	18	4	-7	212	563
Sept	2	5	17	12	1	-13	-25	131	482
Oct	10	10	-13	-11	12	0	-9	9	360
Nov	-21	-9	-28	-6	23	23	15	-107	244
Dec	-29	5	-17	-16	10	21	27	-183	168
Effect	213	114	53	-10	-70	-126	-174	351	000
Mean	564	467	404	341	281	225	177	000	-351

(Cont'd)

exhibit 1 (Cont'd)

C) RESIDUALS AS A BATCH -- same units

|To|		
	-4	
2	-3	44
13	-2	2432265189
27	-1	254815839233176
42	-0	705418475570996
42	0	82267186242515
28	1	645153687200250
13	2	95616423317
2	3	0
1	4	6

42 + 42 = 84, hinges at 21h

Hinges:	-13 and +13h
Midspread:	26h
Sides:	-39h and +40
Outside:	0 and 1
(count/20):	(4)

$$\text{clarity} = \frac{213 + 174 + 227 + 198}{26.5} = 30.8$$

D) ORDERED and GROUPED EFFECTS

Months

-198
-183 } DJF
-181
-122 } MN
-107
-34 A
9 O
71 M
131 S
176 J
212 } JA
227

Places

-174 C
-126 P
-70 B
-10 W
53 A
114 BR
213 L

E) EFFECT-ORDERED RESIDUALS -- still in units of 10 $(°F)^2$

	-174	-126	-70	-10	53	114	213
-198	29	16	2	-12	-7	8	-34
-183	27	21	10	-16	-17	5	-29
-181	14	2	-15	-22	-5	25	0
-122	-8	-1	-18	-14	-4	21	26
-107	15	23	23	-6	-28	-9	-21
-34	-23	-15	-24	-11	11	15	46
9	-9	0	12	-11	-13	10	10
71	-13	-22	-18	7	15	6	26
131	-25	-13	1	12	17	5	2
176	-12	-19	-5	22	24	-7	-4
212	-7	4	18	16	2	-26	-5
227	6	8	13	30	1	-34	-22

(Cont'd)

exhibit 1 (Cont'd)

F) CONDENSED RESIDUALS -- medians of above

	Caribou	Portland	Boston	Wash'n	Atlanta	Baton Rouge	Laredo
DJF	27	16	2	-12	-7	8	-29
MN	4	11	2	-10	-16	6	2
Apr	-23	-15	-24	-11	11	15	46
Oct	-9	0	12	-11	-13	10	10
May	-13	-22	-18	7	15	6	26
Sept	-25	-13	1	12	17	5	2
June	-12	-19	-5	22	24	-7	-4
JA	0	6	16	23	2	-30	-14

G) CONDENSED RESIDUALS AS A BATCH -- same units

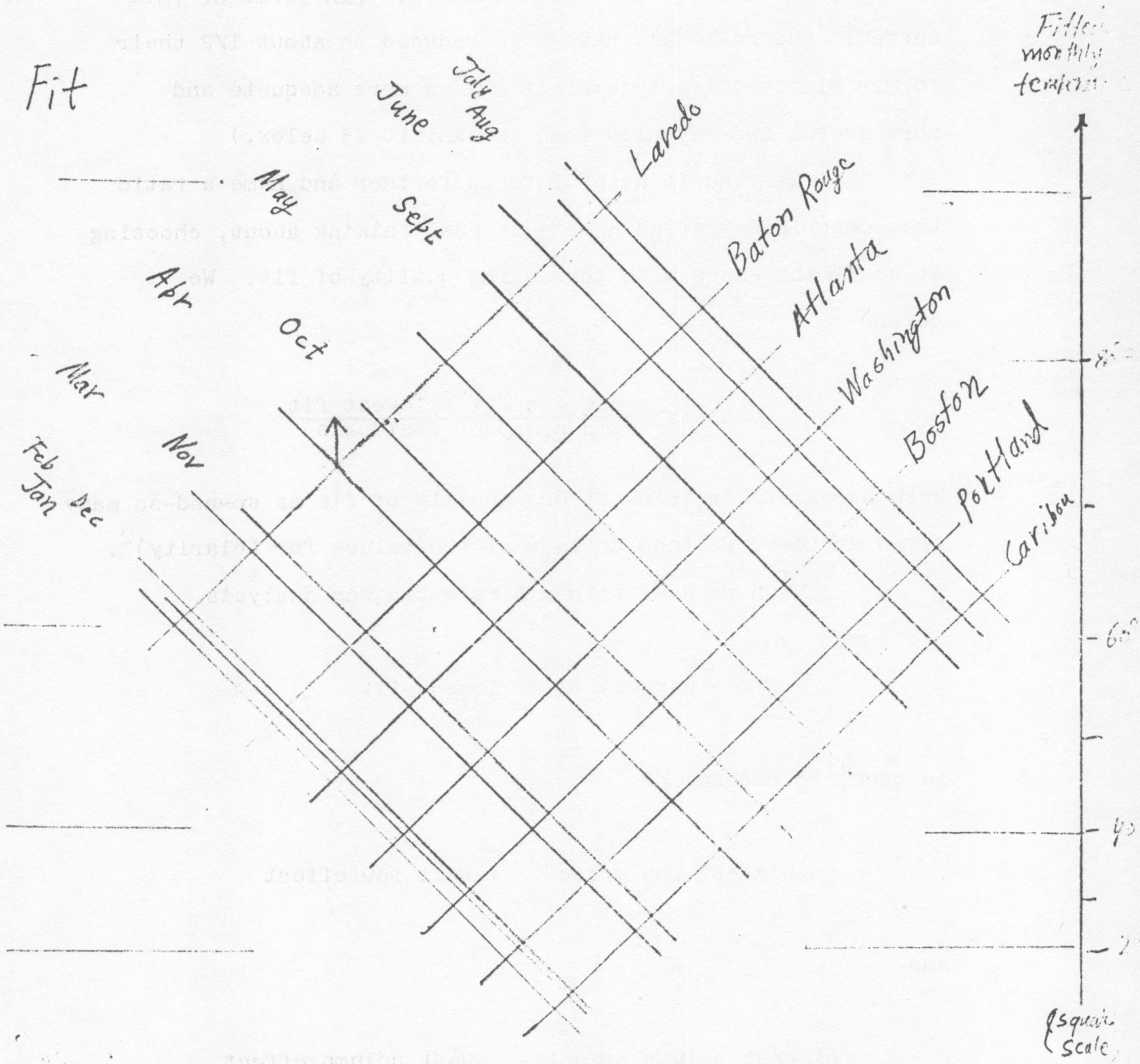
|To|		
	-4	
1	-3	0
6	-2	93425
19	-1	2065113383294
26	-0	7905780
(13)	0	2842627615162
17	1	61152005276
6	2	76243
.	3	
1	4	6

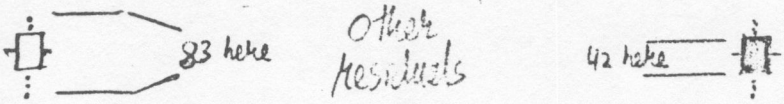
26 + 13 + 17 = 56, hinges at 14h

Hinges:	-12h	and 11
Midspread:		23h
Sides:	-36	and 34h
Outside:	0	and 1
(count/20):		(3)

exhibit 2 of chapter 18

Two-way plot of fitted squared temperatures





of the largest effects to less than 7%. (In terms of (midspread)^2 the residuals have been reduced to about 1/2 their former size. Clearly exhibit 2 is a more adequate and more useful two-way plot than is exhibit 23 below.)

We find it helpful to go further and name a ratio that describes what we have just been talking about, choosing it so it increases with increasing quality of fit. We define

$$\text{clarity} = \frac{\text{highest fit - lowest fit}}{\text{midspread of residuals}}$$

When necessary to think of one quality of fit as so-and-so many times another, we tend to look at the values for (clarity)^2.

When we deal with a simple two-way analysis,

highest fit - lowest fit

is equal to the sum of

highest row effect - lowest row effect

and

highest column effect - lowest column effect.

Since highest effects are never negative, and lowest effects are never positive, this means that -- for simple two-way analyses only --

highest fit - lowest fit

is equal to the sum of 4 terms, namely

the sizes of highest and lowest row and column effects.

Exhibit 3 shows the condensed residuals. There is no trace of a four-alternate corners pattern. However, there are distinct tendencies toward up-and-down coast tilts. April and May -- and perhaps June and September -- appear high at Laredo and low in Maine. July-August and December-January-February appear rather the reverse.

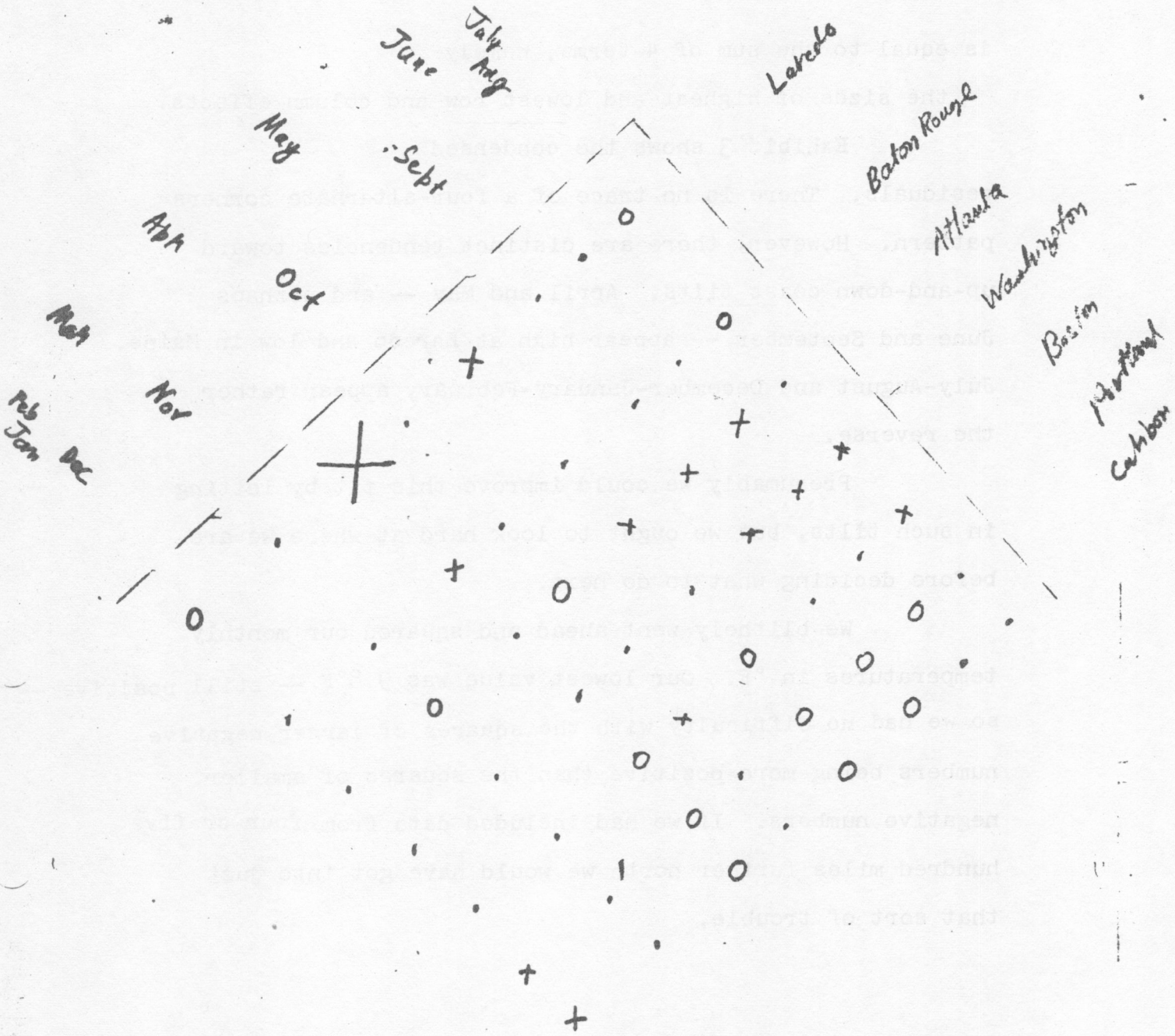
Presumably we could improve this fit by letting in such tilts, but we ought to look hard at where we are before deciding what to do next.

We blithely went ahead and squared our monthly temperatures in °F. Our lowest value was 9.8°F -- still positive -- so we had no difficulty with the squares of larger negative numbers being more positive than the squares of smaller negative numbers. If we had included data from four or five hundred miles further north we would have got into just that sort of trouble.

exhibit 3 of chapter 18

Condensed residuals over a two-way plot

(see exhibit 2)



Moreover, squared temperature could not work for the Arizona monthly temperatures we examined at the start of chapter 16.

We cannot expect squared temperature to be generally useful. The improvement it made in analyzing just this set of data was valuable -- our two-way plot could reasonably be thought to be twice as good -- but this improvement has to be considered a one-shot affair.

Before we add further complexities to an analysis made in such transient terms, we ought to go back and try adding them to the original analysis.

18B. How much change of expression?

If we are going to try a change of expression, it is nice to have guidance about how far to go, as well as in which direction. The four-point approach -- taken at the start of the last section -- does well if the pattern of the residuals is smooth and well-behaved. Since we often have to deal with irregular and misbehaving residuals, we need an alternate approach.

Let us take a quite empirical attitude. Exhibit 4 gives a small table of data where moving from one row to another -- or from one column to another -- corresponds to multiplication by a constant factor -- by some power of 2. If we were to analyze logs, all would be wonderful.

exhibit 4 of chapter 18

An example where logs would help

A) The RAW DATA -- twice for later convenience

one copy				another			
2	4	16	32	2	4	16	32
4	8	32	64	4	8	32	64
16	32	128	256	16	32	128	256

B) TWO-WAY ANALYSIS -- by means and by medians

by means					by medians				
31	25	-7	-49	-36	8	6	-6	-22	-10
19	15	-5	-31	-22	0	0	0	0	0
-49	-41	11	81	58	-48	-36	36	132	60
-43	-35	9	67	50	-16	-12	12	44	20

C) Values of (ROW EFF)(COL EFF)/(COMMON)

(from directly above)				(from directly above)			
31	25	-6	-48	8	6	-6	-22
19	15	-4	-29	0	0	0	0
-50	-41	10	78	-48	-36	36	132

How can we see this from the patterns of residuals when we analyze it raw?

If we pick a row with a substantial row effect, and a column with a substantial column effect, the sign of the residual at their intersection is the product of the signs of the two effects, as the reader may easily verify. Looking at

$$(\text{row effect}) \times (\text{column effect})$$

and comparing the values of our residuals to the value of this product might thus seem reasonable.

This is not good enough. Changing responses from feet to inches would multiply residuals by 12 and this product by $12^2 = 144$. This is inconvenient, to say the least. The simplest way out is to change to

$$\frac{(\text{row effect}) \times (\text{column effect})}{(\text{common term})}$$

which is multiplied by $12^2/12 = 12$ by such a change. We shall call the values of this quantity

comparison values

Panel C of exhibit 4 shows the values of this quantity, which is -- in this example -- hardly different in size and sign from the residuals. This is true whether we analyze by means or by medians. The two analyses are rather different in

this case, but -- so long as we define effects in a matching way -- the relation of residuals to comparison values is clearly the same.

Exhibit 5 shows the plot of residuals against comparison values for the analysis by means. Clearly the points fall close to a straight line. The eye-fitted line gives a slope of 1.01.

The reader is urged to try a variety of other exmaples where logs will help. He will find many of them give slopes approaching 1.0.

Exhibit 6 and 7 do the same for a two-way table where taking square roots will clean things up neatly. Again a good straight line. This time its slope is close to 1/2.

The reader is urged to try other examples which would be rather thoroughly helped by a square-root. He will find many of them give a slope close to 1/2.

We can save arithmetic by avoiding the repeated divisions by "common term". Our purpose is a plot. We lose nothing by calculating -- and plotting on the horizontal scale of our graph paper -- (row eff) × (col eff) and then marking in a new scale for use in assessing slope. When we do this, it pays to note down the largest size of any actual comparison value.

Exhibits 8 and 9 do the same for a two-way table where taking reciprocals will clean things up neatly. This time we have simplified matters somewhat taking 2 half-steps of MEDIAN polish, calculating (col eff) × (row eff), turning the plot the better way, and using the two scales on the plot. This time the slope is 1.77, which we can call close to 2. The reader is encouraged to try various tables where reciprocals are the expression that will help most. He should be a little surprised how many give plots reasonably close to a straight line, with slope somewhere in the vicinity of 2.

exhibit 5 of chapter 18

Residuals vs. comparison values
for an example where logs would help
(see exhibit 4)

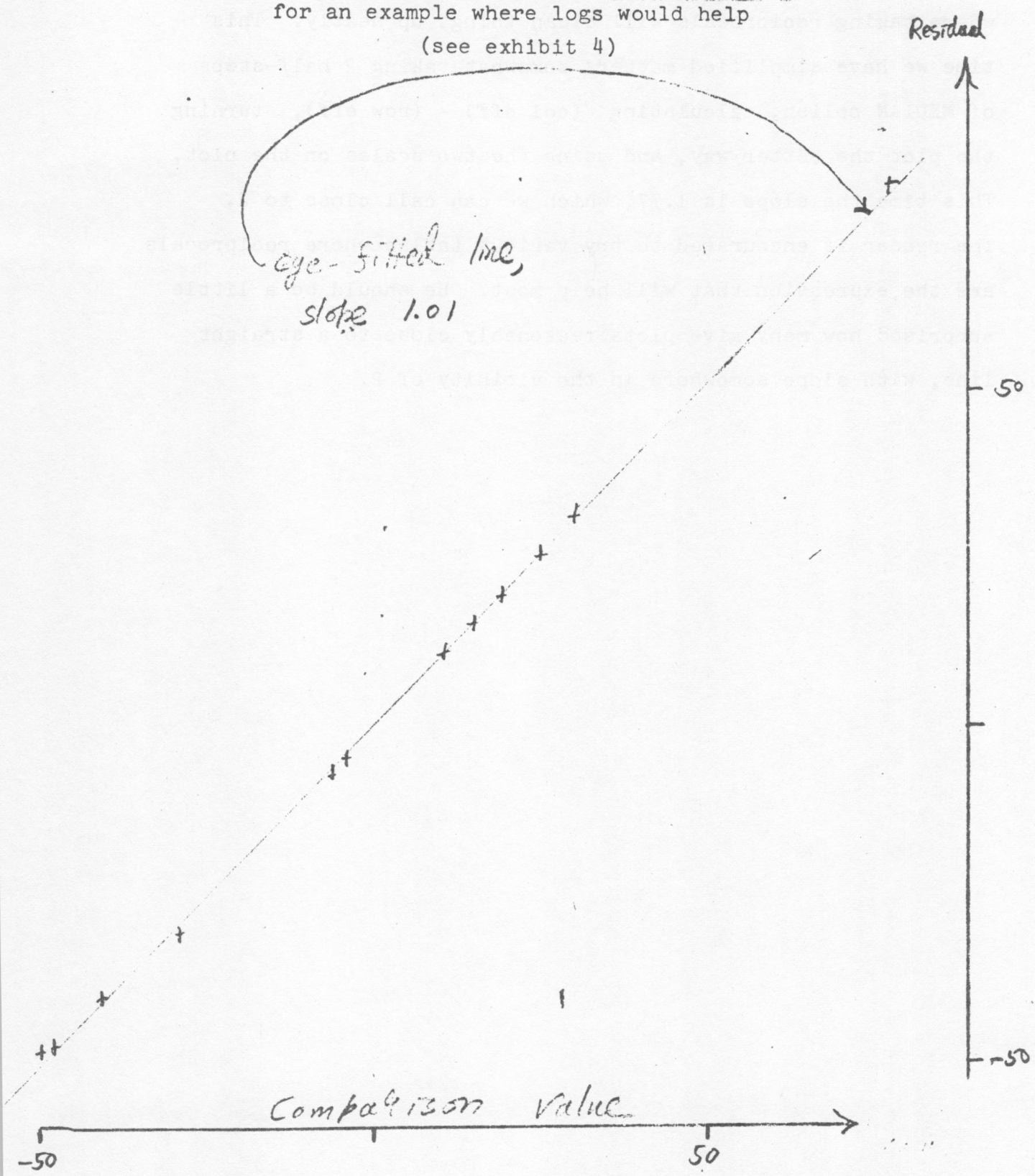


exhibit 6 of chapter 18

An example where roots would help

A) The RAW DATA

9	25	36	36	49
25	49	64	64	81
49	81	100	100	121
81	121	144	144	169

B) The ELEMENTARY ANALYSIS

14.4	2.4	-3.6	-3.6	-9.6	-46.4
4.8	0.8	-1.2	-1.2	-3.2	-20.8
-4.8	-0.8	1.2	1.2	3.2	12.8
-14.4	-2.4	3.6	3.6	9.6	54.4
-36.4	-8.4	8.6	8.6	27.6	77.4

C) Values of (ROW EFF) × (COL EFF)/(COMMON)

22	5	-5	-5	-17
10	2	-2	-2	-7
-6	-1	1	1	5
-26	-6	6	6	19

exhibit 7 of chapter 18

Residuals vs. comparison values for an example where square roots will help (see exhibit 6)

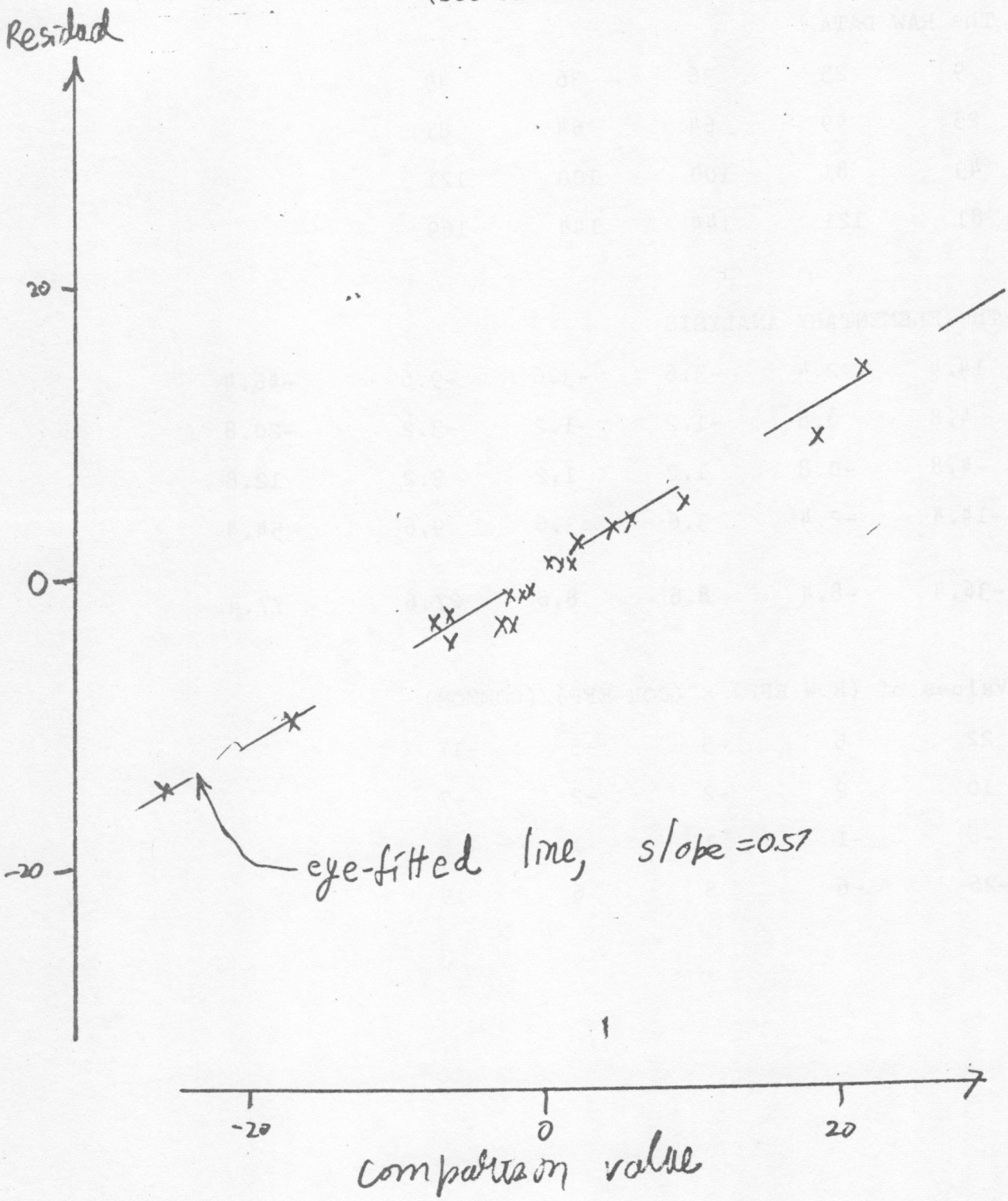


exhibit 8 of chapter 18

An example of a table
where reciprocals would help

A) The TABLE of RECIPROCALS -- 1000/DATA

10	12	16	18	25
11	13	17	19	26
15	17	21	23	30

B) The RAW DATA

100	83	62	56	40
91	77	59	53	38
67	59	48	43	33

C) TWO-WAY ANALYSIS -- by medians; row then columns,
2 half-steps

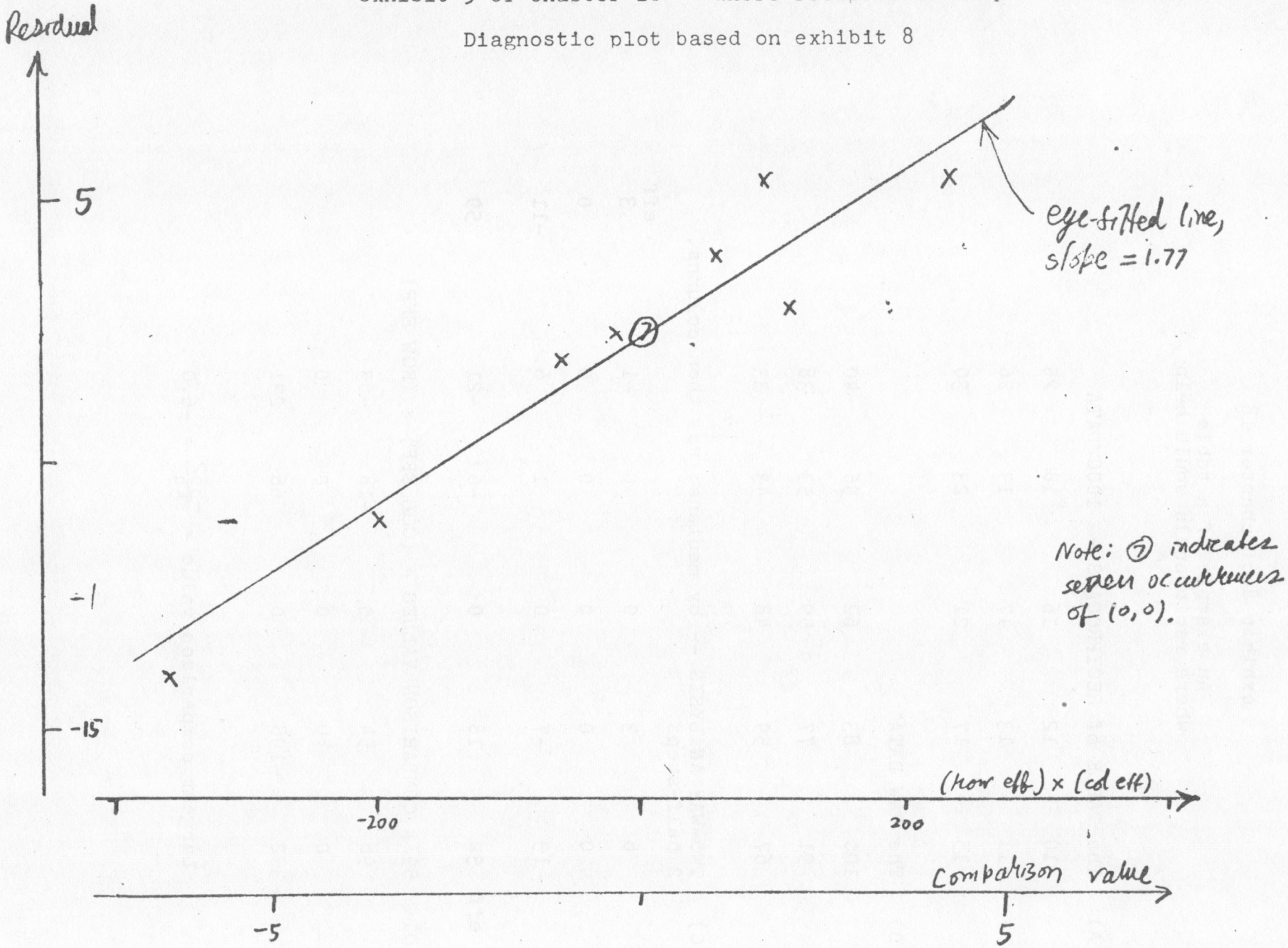
						eff
	6	3	0	0	-1	3
	0	0	0	0	0	0
	-13	-7	0	1	6	-11
eff	32	18	0	-6	-21	59

D) 59 × COMPARISON VALUES = (COL EFF) × (ROW EFF)

96	54	0	-18	-63
0	0	0	0	0
-352	-198	0	66	231

Largest comparison value = $\frac{-352}{59} = -6.0$.

Diagnostic plot based on exhibit 8



Quite roughly then, we have:

Good Expression	Change in Exponent	Slope
Raw	none	near zero
Root	1/2	near 1/2
Log	1 (*)	near 1
Reciprocal	2	near 2

(*) Recall that the log takes the place of the zero power (see Section 3G).

The message is clear. The slope of a line eye-fitted to a plot of

residuals

versus

$$\frac{(\text{row effect}) \times (\text{column effect})}{(\text{common term})}$$

will often be a useful first approximation to the decrease in exponent required to reach a good expression.

We have learned two things in this section:

- how to make a first approximation to the decrease in exponent that would help us most in a simple change of expression.

- that situations where a change in expression would help are also likely to be helped by adding an appropriate multiple of

$$\frac{\text{(row effect)(column effect)}}{\text{(common term)}}$$

to the fit formed by elementary analysis.

We shall illustrate the use of first one and then the other of these two results in the sections that follow.

18C. Seizing a better expression

In this section, and the next, we shall use comparison values as a guide in choosing better expressions.

* catch of plankton *

In 1940, Winsor and Clarke reported the results of a study of the catch of plankton (small ocean life) in nets of different size and in nets of the same kind hauled successively or simultaneously. Exhibit 10 shows a small part of their data. If we proceed to an elementary analysis, using hauls and net-depth combinations as the factors we find the results shown in panel B. The evidence of something large and systematic going on is overwhelming. We need to find out:

- what is going on in this set of data.
- how to deal with it.

exhibit 10 of chapter 16

Catch of Sagitta elegans by Plankton Nets
of 2 Diameters at 3 Depths Hauled Simultaneously
(Many catches estimated from samples.)

A) THE DATA

Haul	12.7 cm at 29 meters (A)	75 cm at 30 meters (B)	12.7 cm at 31 meters (C)
1	45	3020	7
2	100	4350	62
3	117	6060	76
4	110	6600	65
5	(one haul lost)		
6	27	1240	27
7	31	1240	52
8	20	1040	7
9	24	415	6
10	27	690	5

B) ELEMENTARY ANALYSIS - USING MEANS

	29	30	31	eff	fit
1	-90	200	-110	81	1024
2	-515	1050	-535	561	1504
3	-1078	2180	-1101	1141	2084
4	-1259	2546	-1286	1315	2258
5					
6	485	-987	503	-512	431
7	478	-996	519	-503	440
8	553	-1112	558	-587	356
9	765	-1529	765	-795	148
10	675	-1347	671	-702	241
eff	-889	1796	-907	943	000
fit	54	2739	36	000	-943

(Cont'd)

exhibit 10 (Cont'd)

C) Values of (ROW EFF)(COL EFF)/(COMMON) (*)

1	-76	155	-78
2	-540	1080	-510
3	-1080	2180	-1100
4	-1240	2500	-1270
5			
6	480	-980	490
7	470	-960	480
8	550	-1120	570
9	750	-1520	770
10	666	-1340	680

(*) To sliderule accuracy only.

D) RESIDUALS of B as a BATCH -- unit 100

|to|					
1	-1	5	13 + 1 + 13 = 27, hinges at 7		
9	-1	11330013		(cut)	(full)
11	-0	55	Hinges:	-11 and +7	-1078 and 671
13	-0	11	Midspread:	18	1749
14	0	2	Sides:	-29 and +25	
13	0	5555668877	Outside:	0 and 1	
3	1	0	(count/20):	(1h)	
	1				
2	2	26			

$$\text{Clarity} = \frac{(2258 - 148) + (2739 - 36)}{1749} = 2.7$$

$$\text{(or)} \quad = \frac{1315 + 795 + 1796 + 907}{1749} = 2.7$$

E) SOURCE: C. P. Winsor and G. L. Clarke 1940. A statistical study of variation in the catch of plankton nets. Journal of Marine Research 3: 1-34, especially table VIII at page 33.

Data location: Cape Cod Bay, off Canal Entrance; 9 September 1937; 10 minute hauls at 2 knots.

Panel C calculates the values of

$$\frac{\text{(row effect)(column effect)}}{\text{(common term)}}$$

A quick look shows that these very nearly duplicate the residuals. The plot of residuals against these quantities must, therefore, be very nearly a straight line with slope 1.00. Thus we are firmly told to try logs.

What is going on is now, I hope, clear to all of us who have stopped to look at exhibit 9 and think about its numbers. The big net, six times the diameter of the small nets, would naively be expected to catch something like 36 times as much as the small nets. While the catch in the big net is falling by 5000, say from 6000 to 1000, we would expect a proportionate fall in the catch in the small nets, a fall by one or two hundred -- certainly not a fall by 5000 from 200 to an impossible figure of MINUS 4800. Differences in numbers caught by nets of different size are just not a sensible thing to look at. We should be looking at ratios. As usual, the easy way to look at ratios is to look at differences of logs.

Exhibit 11 shows the result. The figure of clarity has risen from 1.7 (quite poor) to about 10. This means something like 30 times the information. (A simple resistant analysis would raise it to 17.) Clearly logs

exhibit 11 of chapter 16

Analysis of log plankton counts
(see exhibit 10)

A) The DATA

	31	29	30
4	1.81	2.04	3.82
3	1.88	2.07	3.78
2	1.79	2.00	3.64
1	.85	1.65	3.48
6	1.43	1.43	3.09
7	1.72	1.49	3.09
8	.85	1.30	3.02
9	.78	1.38	2.62
10	.70	1.43	2.84

B) Its ELEMENTARY ANALYSIS

	31	29	30	Eff	Mean
4	.00	-.08	.08	.48	2.56
3	.05	-.07	.02	.50	2.58
2	.06	-.04	-.02	.40	2.48
1	-.40	.09	.30	-.08	2.00
6	.19	-.12	-.08	-.09	-1.99
7	.37	-.17	-.19	.02	2.10
8	-.13	.01	.11	-.35	-1.73
9	-.07	.22	-.16	-.48	-1.60
10	-.21	.21	.00	-.42	-1.66
Effect	-.75	-.44	1.18	2.08	.00
Mean	1.33	1.64	3.26	.00	-2.08

C) The RESIDUALS as a BATCH -- unit 0.01

1	-2	1
4	-1	236
13	-0	874249817
(9)	0	085261110
5	1	291
2	2	21

13 + 9 + 5 = 27, hinges at 7

Hinges:	-.08 and +.06
Midspread:	.14
Sides:	-.22 and +.20
Outside:	0 and 2
(count/20):	(1h)

Extreme effects: +.50, -.48, +1.18, -.75

$$\text{Clarity} = \frac{.50 + .48 + 1.18 + .75}{.14} = 10$$

(Cont'd)

exhibit 11 (Cont'd)

D) The EFFECTS CONDENSED

Haul	Effect	Split	Mark	Net	Effect	Split	Mark
3	.50			31	-.75		31
4	.48		342	29	-.44		29
2	.40			30	1.18		30
7	.02						
1	-.08		716				
6	-.09						
8	-.35						
10	-.42		8109				
9	-.48						

E) The RESIDUALS CONDENSED

	31		29		30		
3	.05		-.07		.02		-1 | 23
4	.00	.05	-.08	-.08	.08	.02	-0 | 88
2	.06		-.04		-.02		0 | 520
							1 | 9
7	.37		-.17		-.19		2 | 1
1	.00	.19	-.08	-.12	.08	-.08	
6	.19		-.12		-.08		
8	-.13		.01		.11		
10	-.21	-.13	.21	.21	.00	.00	
9	-.07		.22		-.16		

hinges at depth 3

Hinges:	-.08	and	.05
Midspread:		.13	
Sides:	-.21	and	.18
Outside:	0	and	2
(count/20):		(h)	

were very much in order. Since we now have an analysis worth looking at, exhibit 12 gives the two-way plots of fitted values and residuals. Nothing of any easy interpretation appears in the plot of condensed residuals, so we may feel that we have probably finished the analysis.

Naive computation suggested that the large net ought to catch about $6^2 = 36$ times as many Sagitta. Since log 36 = 1.56, we might try reducing the log catch for the large net by 1.56. When we do, we get the following mean logs

29 meters (small net)	1.64
30 meters (large adj.)	1.70
31 meters (small net)	1.33

The agreement is reasonable, though there is a suggestion that the large net may be catching more than 36 times the catch of a small net.

18D. Feeling around for a better expression

* control of insects *

Let us next turn to data of Geoffrey Beall on the control of numbers of insects, here Leptinotarsa decemlineata. Each of 7 blocks was divided into 4 plots, and four different treatments applied one to a plot. Exhibit 13 gives the data and its initial analysis, using 2 median half-steps, and rounding the effects thus found.

Exhibit 14 gives the diagnostic plot of residuals against the product of ROUNDED effects. (We only need a rough slope, so we may as well save arithmetic.) The eye-fitted slope of 0.8, nearer to 1 than to .5, suggests trying logs.

Two-way plots for fitted values and coded condensed residuals

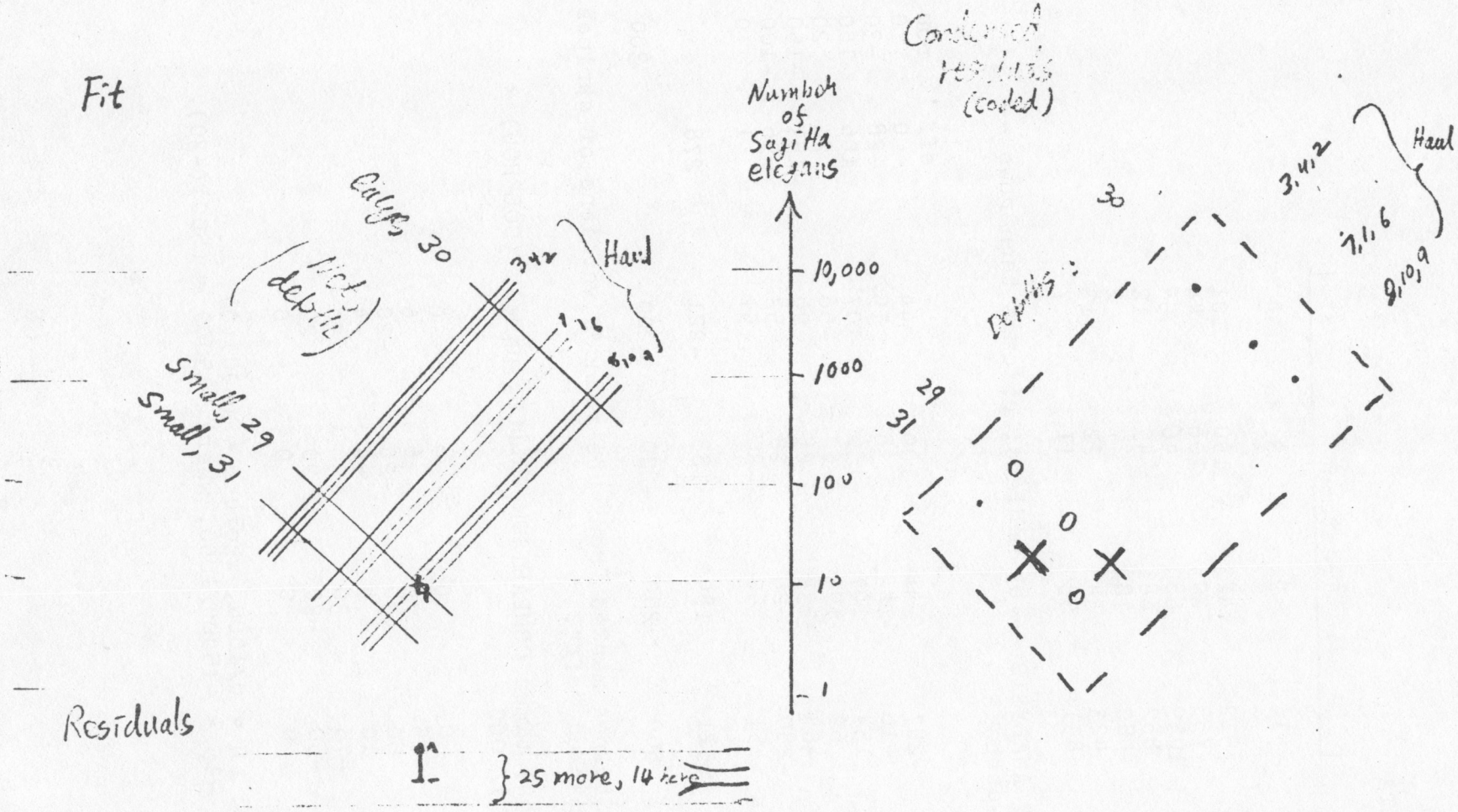


exhibit 13 of chapter 18

Counts per plot of Leptinotarsa Decemlineata

A) The DATA

Block	Treatment 1	2	3	4
1	512	423	190	101
2	755	295	100	44
3	1059	680	46	56
4	882	681	49	12
5	580	182	7	2
6	423	343	22	13
7	851	530	11	7

B) TWO-WAY ANALYSIS -- by medians; columns then rows -- 2 half-steps

					eff	rou
1	-287	-44	100	44	0	0
2	-16	-144	38	15	-28	-30
3	154	107	-150	-107	106	100
4	62	193	-62	-66	21	20
5	-68	-134	68	96	-151	-150
6	-280	-28	28	52	-96	-100
7	51	62	-80	-51	1	0
eff	521	189	-188	-221	278	
rou	500	200	-200	-200		300

Note: Stripes marked "rou" are rounded versions of stripes marked "eff".

C) (300/1000) times COMPARISON VALUES, namely (COL ROU) × (ROW ROU)/1000

1	0	0	0	0
2	-15	-6	6	6
3	50	20	-20	-20
4	10	4	-4	-4
5	-75	-30	30	30
6	-50	-20	20	20
7	0	0	0	0

Examples: 0 = 0/1000, where 0 = (500)(0),
-15 = -15000/1000, where -15000 = (500)(-30).

exhibit 13 (cont'd.)

D) FURTHER POLISHED ANALYSIS -- column medians, row medians -- total 4 half-steps

					eff
1	-277	-22	66	23	5
2	0	-116	10	0	-29
3	164	129	-184	-128	111
4	80	223	-88	-79	18
5	-62	-140	62	59	-174
6	-264	0	0	37	-97
7	67	90	-108	-66	0
eff	506	161	-160	-206	279

E) RESIDUALS as a BATCH -- unit 10, cut

to		
	2	
1	2	2
2	1	6
3	1	2
9	0	869665
(5)	0	02123
14	-0	000
11	-0	6876
7	-1	1402
3	-1	8
1	-2	
2	-2	76

9 + 5 + 14 = 28, hinges at 7h

	(cut)	(restored)
Hinges:	+ 6 and -9	+ 64 and - 98
Midspread:	15	162
Sides:	+21 and -24	+222 and -260
Outside:	1 and 2	1 and 2
(count/20):	(1h)	(1h)

Clarity

$$\text{Clarity} = \frac{506 + 206 + 111 + 174}{162} = 6.1$$

F) SOURCE

Geoffrey Beall 1942. The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. Biometrika 32: 243-262 calculated from table 3 on page 245.
Note: Leptinotarsa decemlineata is the Colorado potato beetle.

exhibit 14 of chapter 18 -- control of beetles

ex bit 14 of chapter 18 -- control of beetl

Diagnostic Plot
Counts of Leptinctarsa decemlineata

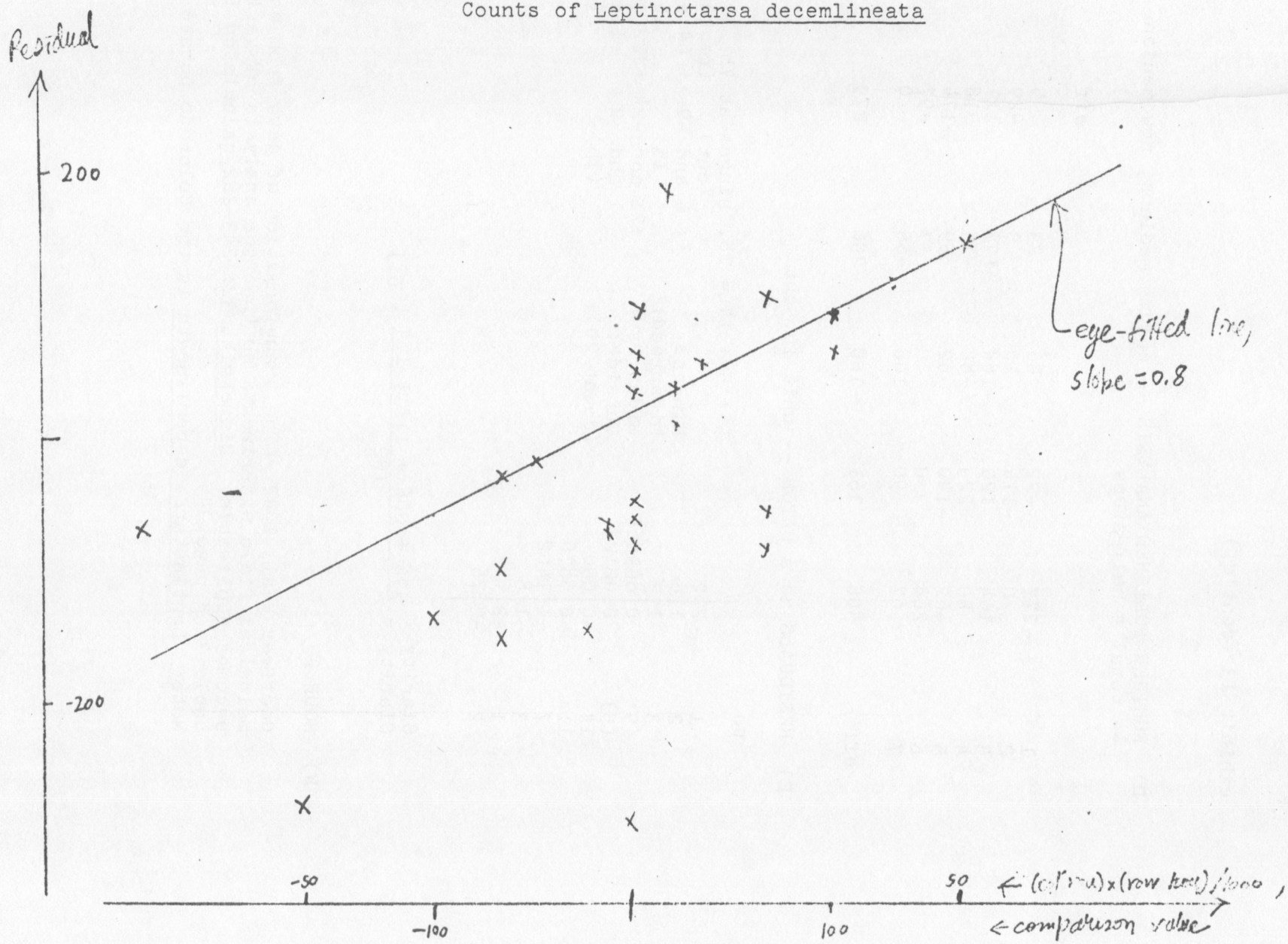


Exhibit 15 gives an analysis of log numbers of beetles. Exhibit 16 shows the residuals against the comparison values. The slope is -2.0. If we took this seriously, we would go next to the third power of the expression used here, namely to

$$(\text{log of count})^3$$

We do not think this very reasonable.

Let us look around a little. For a few values spread across the counts in the data we have:

count	log count	$(\text{log count})^3$	$\sqrt{\text{count}}$
1000	3.0	27	32
100	2.0	8	10
10	1.0	1	3
1	0.0	0	1

Certainly exhibit 14 said "lower than raw count", while exhibit 16 said "higher than log count". Square-root of count now looks good. It should be the natural next trial.

Exhibit 17 analyzes the square-roots of count. Exhibit 18 plots residuals against comparison values. No clear tendency to tilt appears to the naked eye. A reasonable analyst will now be happy with square roots.

Exhibit 19 displays the squre-root fit. Treatments 3 and 4 seem to have considerably reduced the number of beetles.

exhibit 15 of chapter 18 -- control of beetles

Analysis of log counts
(see exhibits 12, 13)

A) The DATA

Block	Treatment 1	2	3	4
1	2.71	2.63	2.28	2.00
2	2.88	2.47	2.00	1.65
3	3.02	2.83	1.66	1.75
4	2.95	2.83	1.69	1.08
5	2.76	2.26	.85	.30
6	2.63	2.54	.34	1.11
7	2.93	2.72	.04	.85

B) Its TWO-WAY ANALYSIS -- by medians; cols then rows, 2 half-steps

					eff
1	-.48	-.31	.31	.58	.31
2	-.17	-.33	.17	.37	.17
3	-.03	.03	-.17	.47	.17
4	.02	.15	-.02	-.08	.05
5	.47	.22	-.22	-.22	-.59
6	-.08	.08	-1.15	.17	-.17
7	.15	.19	-1.52	-.16	-.10
eff	.74	.49	-.48	-1.03	2.14
rou	.7	.5	-.5	-1.0	

C) 2.14 × COMPARISON VALUES = (COL ROU) × (ROW EFF)

1	.217	.155	-.155	-.310
2	.119	.085	-.085	-.170
3	.119	.085	-.085	-.170
4	.035	.025	-.025	-.050
5	-.413	-.295	.295	.590
6	-.119	-.085	.085	.170
7	-.070	-.050	.050	.100

Largest comparison value = $\frac{.590}{2.14}$ = .28

(Cont'd)

exhibit 15 (cont'd.)

D) The ANALYSIS of panel B repolished -- cols and row medians, total 4 half-steps

1	-.46	-.40	.47	.40	.29
2	-.17	-.44	.31	.17	.07
3	.00	-.05	.00	.30	.06
4	-.01	.01	.09	-.31	.00
5	.46	.10	-.09	-.39	-.66
6	-.05	.00	-.18	.00	-.28
7	.29	.22	-1.24	-.22	-.35
eff	.75	.61	-.61	-.82	2.21

E) The RESIDUALS as a BATCH

Low	to			
-.98	3	4	670	9 + 5 + 5 + 9 = 28, hinges at 7h
-1.24	5	3	10	
	7	2	92	Hinges: +.19h and -.26h
	9	1	07	Midspread: .46
	(5)	0	01009	Sides: +.65h and -.72h
	(5)	-0	15590	Outside: 0 and 2
	9	-1	7	(count/20): (1h)
	8	-2	2	
	7	-3	19	
	5	-4	604	
	2	L	LL	

Clarity: $\frac{.75 + .82 + .29 + .66}{.46} = 5.5$

ex bit 16 of chapter 18 -- control of beetl

Diagnostic plot for logs of counts

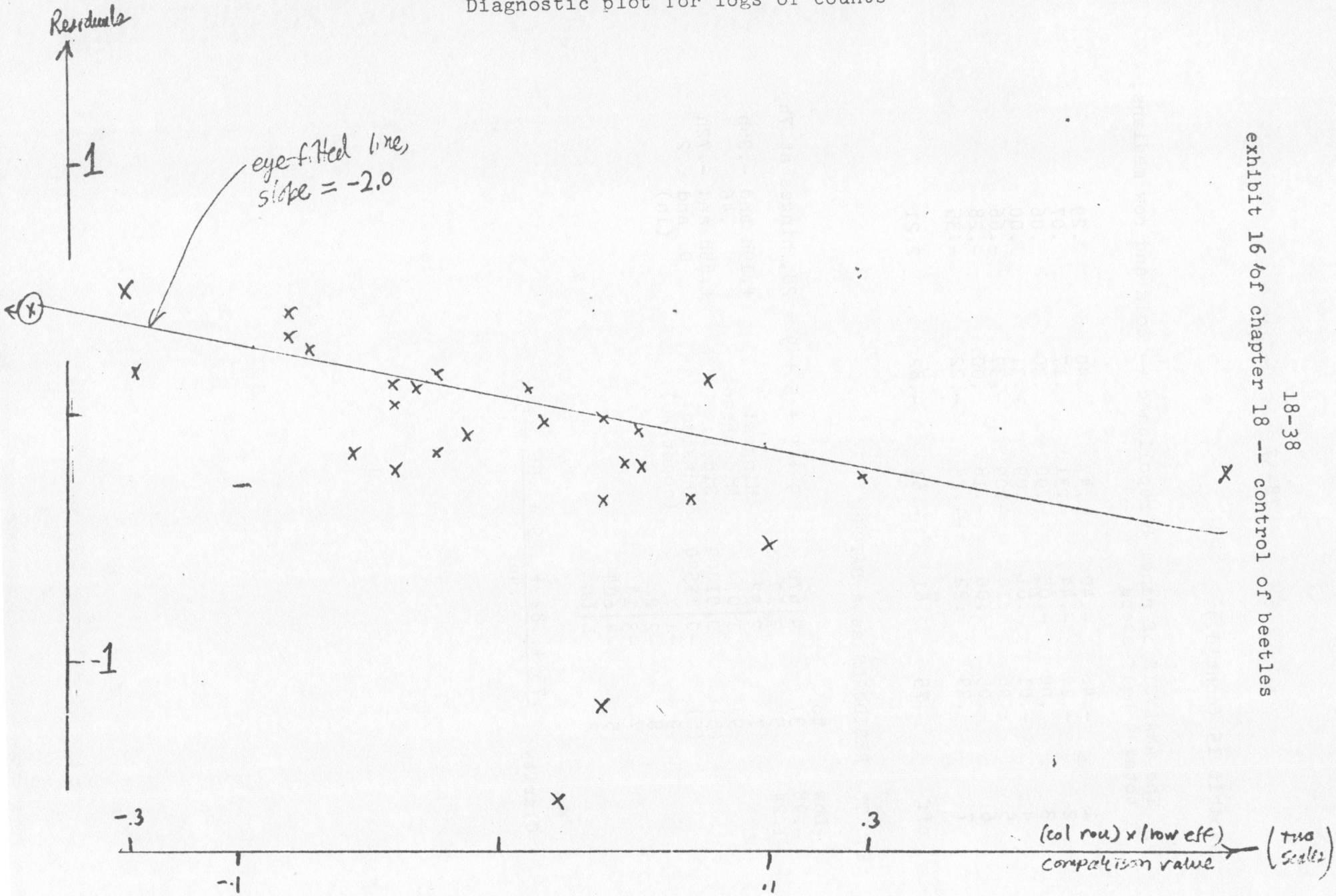


exhibit 17 of chapter 18

Analysis of root number of insects
(see exhibits 13, 14, 15, 16)

A) The DATA

Block	Treatment			
	1	2	3	4
1	23	20	14	10
2	28	17	10	7
3	33	26	7	7
4	30	26	7	3
5	24	14	3	1
6	20	18	5	4
7	29	23	3	3

B) One of its ELEMENTARY ANALYSES -- by medians; stripe medians close to zero

					eff	fit
1	-8	-3	4	3	3	17
2	-2	-5	1	1	2	16
3	1	2	-4	-1	4	18
4	1	5	-1	-2	1	15
5	0	-2	0	1	-4	10
6	-6	0	0	2	-2	12
7	1	3	-4	-1	0	14
eff	14	6	-7	-10	14	00
fit	28	20	7	4	00	-14

C) RESIDUALS as a BATCH

to		
1	5	5
2	4	4
4	3	33
6	2	22
12	1	111111
(4)	0	0000
12	-1	111
9	-2	222
6	-3	3
5	-4	44
3	-5	5
2	-6	6
	-7	
1	-8	8

12 + 4 + 12 = 28, hinges at 7h

Hinges:	+1 and -2
Midspread:	3
Sides:	+4 and -5
Outside:	1 and 2
(count/20):	(1h)

$$\text{clarity} = \frac{3 + 4 + 14 + 10}{3} = 10.3$$

exhibit 17 (cont'd.)

D) COMPARISON VALUES

	1	2	3	4
1	3.0	1.3	-1.5	-2.1
2	2.0	.9	-1.0	-1.5
3	4.0	1.7	-2.0	-2.9
4	1.0	.4	-.5	-1.4
5	-4.0	-1.7	2.0	2.9
6	-2.0	-.9	1.0	1.5
7	.0	.0	.0	.0

E) The EFFECTS CONDENSED -- by eye and rule of thumb

Blocks	#	Treatments	#
4		14	1
3	13		
2		6	2
1	247		
0		-7	
-2			34
-4	56	-10	

F) The RESIDUALS CONDENSED -- medians taken

Blocks	Treatments (condensed) 1	2	34		
13	-4	0	1	Hinges:	0 and -1
				Midspread:	1
247	1	3	-1	Sides:	+1 and -2
				Outside:	1 and 2
56	-3	-1	0	(count/20):	(h)

(cont'd.)

exhibit 17 (cont'd.)

G) The COMPARISON VALUES CONDENSED -- medians taken

	1	2	34
13	3.5	1.5	-2.0
247	1.0	.4	-0.8
56	-3.0	-1.3	1.8

H) PROBLEMS

1) Plot condensed residuals against condensed comparison values. Can you see more -- or less -- than from the plot of individual residuals against individual comparison values? Would you recommend such a plot for large tables, where substantial plotting effort would be saved?

2) Compute a different set of condensed comparison values from condensed effects (and original common value). Do you think these are about as good as those in panel G? Would you recommend their use in large tables where substantial arithmetic would be saved?

I) NOTE

Since the block effects show no clear structure, we have used the "rule of thumb" of chapter 10 to divide them into three -- just as if this were a y vs. x problem.

exhibit 18 of chapter 18

Residuals against comparison values
Root number of beetles

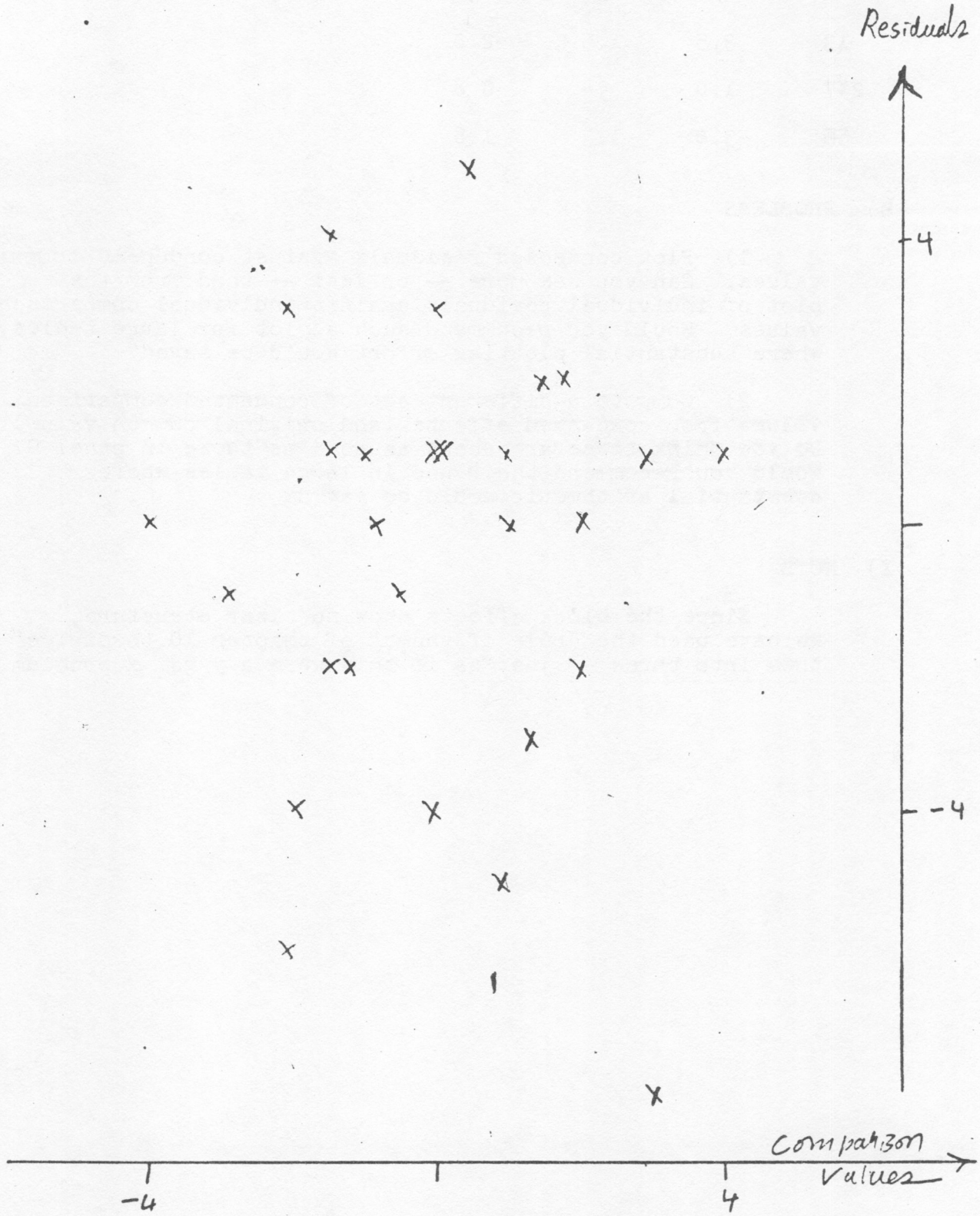
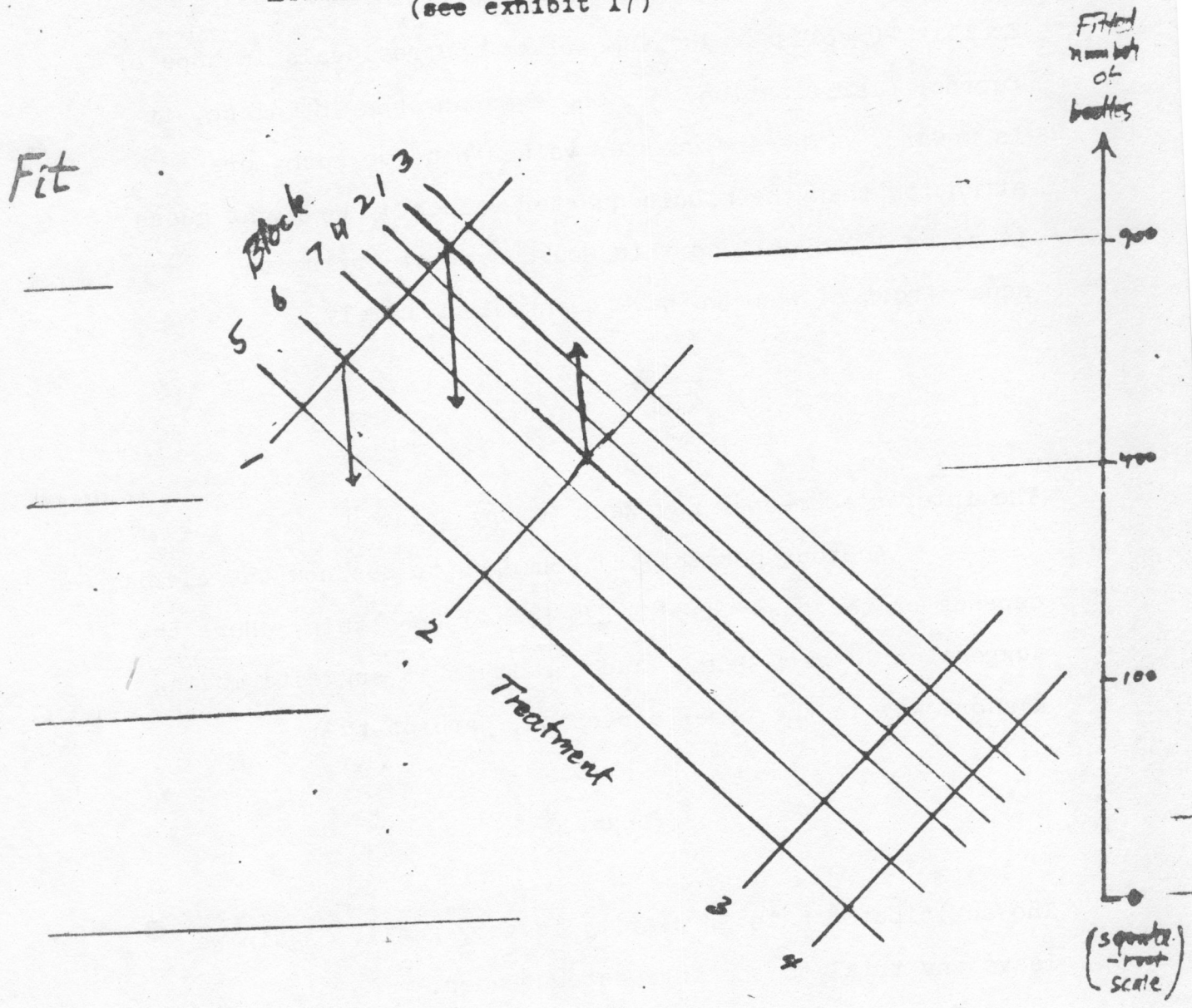
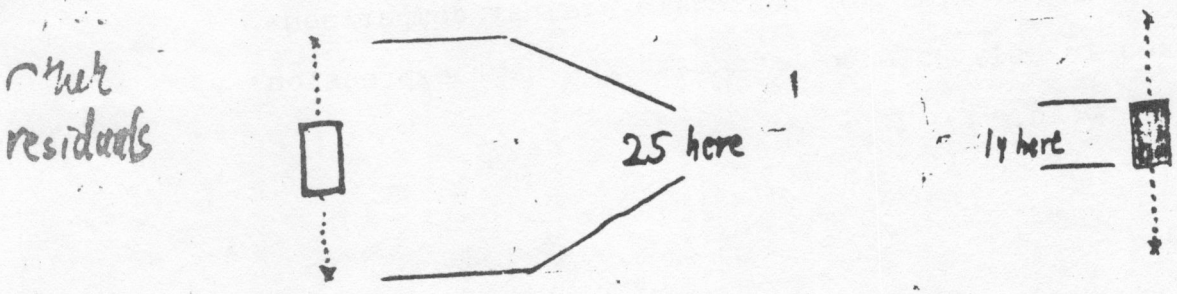


Exhibit 19 of chapter 18

Two-way plot of fitted numbers of beetles
(fitted to square roots)
(see exhibit 17)





* trying to squeeze closer *

Exhibit 20 collects and summarizes the residuals in hope of further clarification. If there is an apparent slope, it is upward. (The end quarter values deserve much more attention than the middle quarter values.) We might guess it to be + 0.5. If so this would suggest trying the square-root of what we have been using, namely

$$\sqrt{\sqrt{\text{count}}} = \sqrt[4]{\text{count}}$$

The interested reader may want to try this.

Another possibility would be to see how the clarity depends on the power used. Exhibit 21 does this. Here the suggestion is very weak. Such as it is it suggests moving a short way in the other direction, perhaps to

$$(\text{count})^{0.6}$$

The anticipated gain in clarity is very small. Again we leave any trial to the interested reader.

There is one thing that the varied suggestions of this example should have made clear:

- why the plot of residuals against comparison values is only said to "suggest" the next expression to be tried.

exhibit 20 of chapter 18

Summary points for exhibit 18

Range of Comparison Values	Values of residuals	Median	Mid-Mean	Tri-Mean
-3.4 to -1.1	-2.8,-1.6,-3.9,1.8,-2.2,-3.5,0.5	-2.2	-2.2	-1.8
-1.0 to -0.1	6.2,3.4,2.9,1.1,-1.1,2.8,2.8	2.8	2.8	2.5
0.1 to 1.2	-6.9,-1.9,0.8,-5.2,3.8,-3.1,-1.9	-1.9	-2.3	-2.0
1.4 to 2.8	3.2,2.1,1.8,0.9,0.1,1.6,1.8	1.8	1.7	1.6

No clear trend. End quarters give slope ~ 1, middle quarters give slope ~ -4.

exhibit 21 of chapter 18 -- control of beetles

comparison of clarities

Expression		Clarity
name	power	
Raw	1.0	6.1
Root	0.5	10.3
Log	.0	5.5

18E. Extension of the fit

In this section we will use comparison values as a way to make a slightly more complicated and slightly more flexible fit. As our example will show, we can often make surprisingly great gains by this simple device.

* East Coast climate *

Let us begin by turning again to the East Coast monthly temperatures. A direct analysis -- see exhibits 22 and 23 -- reaches a clarity of about 20. An analysis of squared temperatures in °F -- in section 18A, there pointed out as a silly but surprisingly useful approach -- showed a clarity of just over 30.

Our task now is to learn how well we can do, while keeping the response in °F, with a fit of the form

$$(\text{common}) + (\text{row}) + (\text{col}) + (\text{constant}')\frac{(\text{row})(\text{col})}{(\text{common})}$$

where

(common) = common term of simple elementary analysis
(row) = row effect " " " "
(column) = column effect of " " "

Mean Monthly Temperatures at 7 East Coast Airports

A) The DATA

	Laredo	Baton Rouge	Atlanta	Wash'n	Boston	Portland	Caribou
Jan	57.6	52.4	44.6	36.2	29.1	20.7	8.7
Feb	61.9	55.6	46.7	37.1	29.2	21.5	9.8
Mar	68.4	60.3	52.7	45.3	37.6	31.9	21.7
Apr	75.9	66.8	61.7	54.4	47.2	41.9	34.7
May	81.2	73.6	70.0	64.7	57.8	52.3	48.5
June	85.8	79.6	77.7	73.4	67.2	61.8	58.4
July	87.7	81.1	79.5	77.3	72.2	67.8	64.0
Aug	87.8	80.7	78.6	75.4	71.5	66.4	61.8
Sept	83.5	77.5	74.4	69.6	64.3	58.6	53.2
Oct	76.5	69.6	63.4	58.2	55.0	48.4	42.1
Nov	66.0	59.1	51.9	47.7	44.4	37.5	29.2
Dec	59.3	53.6	45.2	38.0	32.8	25.1	14.6

B) An ELEMENTARY ANALYSIS

	Laredo	Baton Rouge	Atlanta	Wash'n	Boston	Portland	Caribou	Effect	Temp.
Jan	3.8	5.4	2.9	.3	-1.1	-3.3	-8.0	-20.5	35.6
Feb	6.3	6.8	3.2	- .6	-2.8	-4.3	-8.7	-18.7	37.4
Mar	4.8	3.5	1.2	- .4	-2.4	-1.9	-4.9	-10.7	45.4
Apr	3.0	.7	.9	- .6	-2.1	-1.2	-1.1	-1.4	54.7
May	-1.0	-1.8	- .1	- .4	- .8	- .1	3.4	7.9	64.0
June	-4.4	-3.8	- .4	1.1	.6	1.4	5.3	15.9	72.0
July	-6.2	-6.0	-2.3	1.3	1.9	3.7	7.2	19.6	75.7
Aug	-5.0	-5.3	-2.1	.5	2.3	3.4	6.1	18.5	74.6
Sept	-3.4	-2.6	- .4	.6	1.0	1.3	3.4	12.6	68.7
Oct	- .7	- .8	-1.7	-1.1	1.4	1.0	2.0	2.9	59.0
Nov	-6.2	- .3	-2.2	- .6	1.8	.9	.1	-8.1	48.0
Dec	2.9	3.8	.7	- .7	- .2	-1.9	-4.9	-17.7	38.4
Effect	18,2	11.4	6.1	0.3	-5.4	-11.6	-18.9	56.1	zero
Temp.	74.3	67.5	62.2	56.4	50.7	44.5	37.2	zero	-56.1

(cont'd.)

exhibit 22 (cont'd.)

C) The RESIDUALS as a BATCH

|to|		
1	7	2
4	6	381
6	5	43
7	4	8
16	3	805827444
20	2	9930
30	1	2139048430
39	0	797345691
(16)	-0	7283144646678621
29	-1	087119291
20	-2	6312841
13	-3	483
10	-4	4399
8	-5	03
4	-6	20
	-7	
2	-8	07

39 + 16 + 29 = 84, hinges at 21h

Hinges:	+1.8h	and -1.9
Midspread:		3.7h
Sides:	+5.6	and -5.6h
Outside:	5	and 4
(count/20):		(4)

$$\text{Clarity} = \frac{20.5+19.6+18.4+18.9}{3.75} = 20.6$$

D) SOURCES

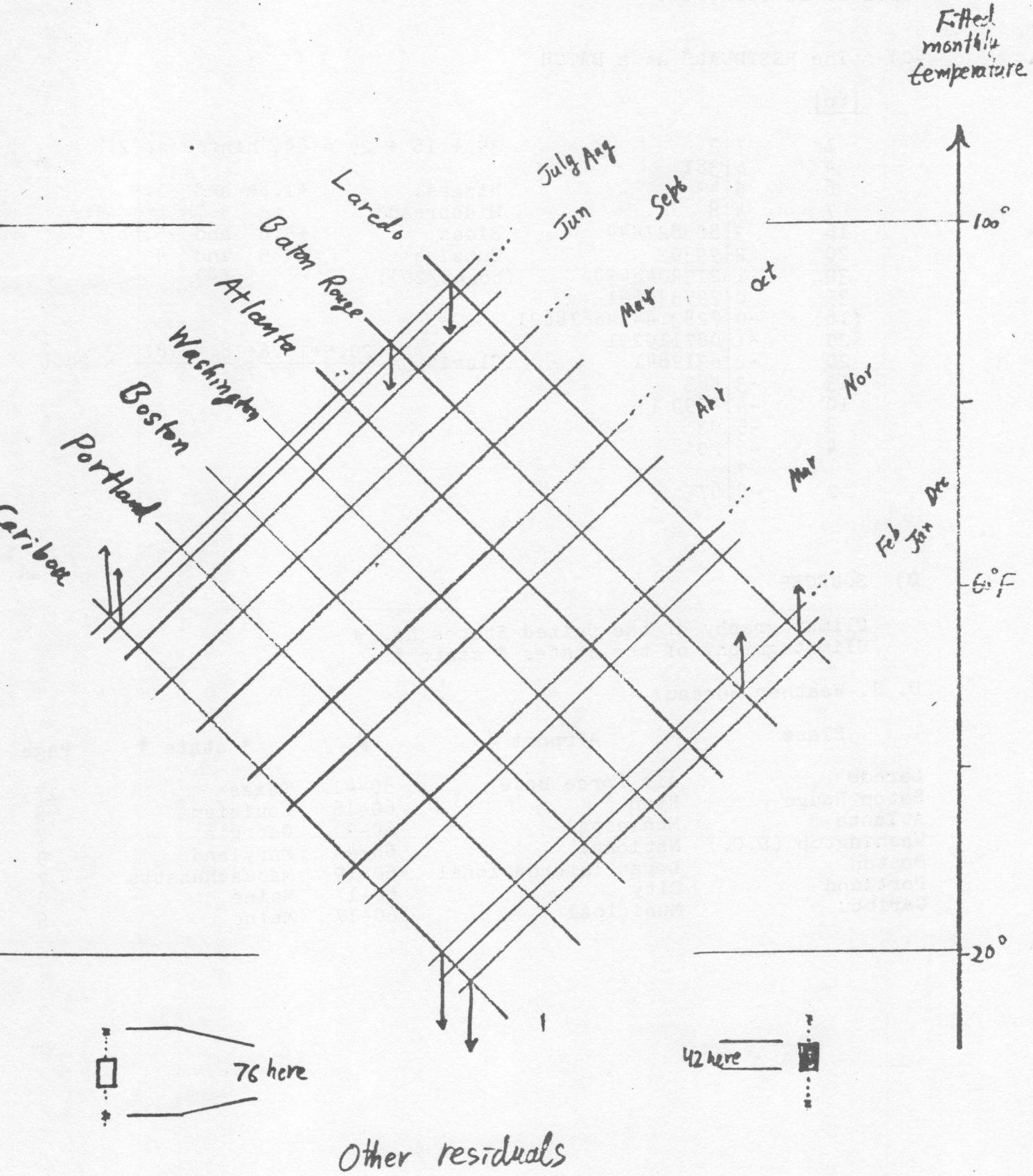
Climatography of the United States No. #
Climatography of the states * state *

U. S. Weather Bureau

Place	Airport	#	* state *	Page
Laredo	Air Force Base	60-41	Texas	17
Baton Rouge	Ryan	60-16	Louisiana	8
Atlanta	Municipal	60-9	Georgia	7
Washington (D.C.)	National	60-44	Maryland	9
Boston	Logan International	60-19	Massachusetts	7
Portland	City	60-17	Maine	8
Caribou	Municipal	60-17	Maine	8

exhibit 23 of chapter 18

Two-way plot of fitted values
East Coast climate



An equivalent form will serve us better, namely

$$(\text{common}) + (\text{row}) + (\text{col}) + \frac{(\text{row})(\text{col})}{A}$$

where A is a constant expressed in the same units as the response. This new constant turns out to be more closely related to our pictures.

Exhibit 24 shows a plot of residuals against comparison values. The eye-fitted slope is -1.05, that of a previous sketch was -0.97. We can do quite well by taking constant' = -1.00 which means taking A = -56.1°F. (If we thought -0.97 was better, we would take A = - 56.1/.97 = - 57.8°F. If we thought -1.05 was better, we would take A = - 56.1/1.05 = - 53.3°F.)

Exhibit 25 gives the comparison values, and the new residuals which, since the fit has been changed by

$$-1.00\ \frac{(\text{row effect})(\text{col effect})}{(\text{common term})} = -\ (\text{comparison value})$$

we can calculate as

(old residuals) + (comparison value)

The midspread of the new residuals is 1.8h, in contrast to the 3.7h of the analysis of exhibit 22.

exhibit 24 of chapter 18

Residuals vs. Comparison Values or

East Coast Temperatures

(see exhibit 25, below, and exhibit 22)

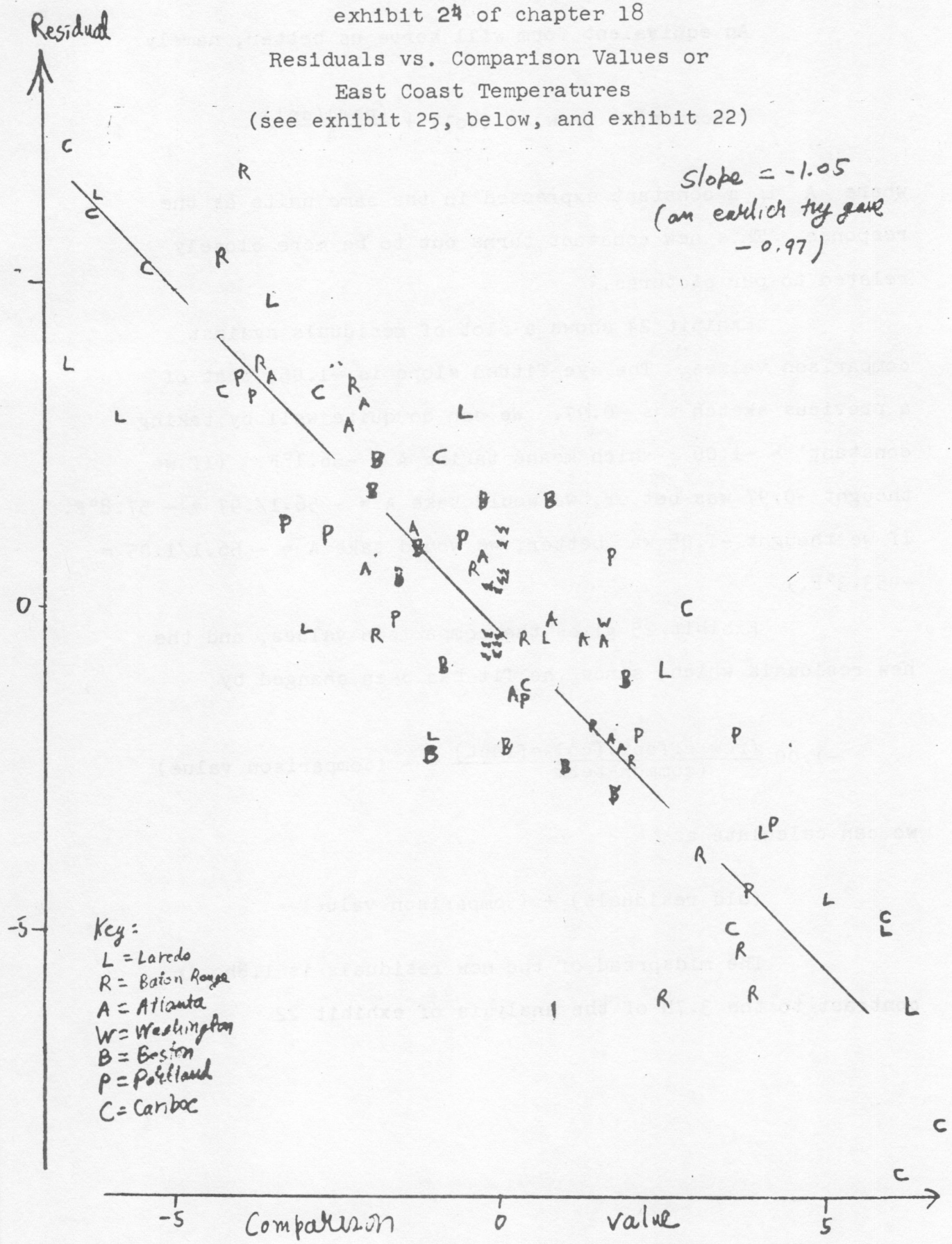


exhibit 25 of chapter 18

Comparison Values, New Residuals, etc.

East Coast Climate

(see exhibit 22)

A) COMPARISON VALUES -- effects in exhibit 22

	Laredo	Baton Rouge	Atlanta	Wash'n	Boston	Portland	Caribou
Jan	-6.7	-4.2	-2.2	-0.1	2.0	4.2	6.9
Feb	-6.1	-3.8	-2.0	-0.1	1.8	3.9	6.3
Mar	-3.5	-2.2	-1.2	-0.1	1.0	2.2	3.6
Apr	-0.5	-0.3	-0.2	0	0.1	0.3	0.5
May	2.6	1.6	0.9	0.1	-0.8	-1.6	-2.7
June	5.2	3.2	1.7	0.1	-1.5	-3.3	-5.4
July	6.4	4.0	2.1	0.1	-1.9	-4.0	-6.6
Aug	6.0	3.8	2.0	0.1	-1.8	-3.8	-6.2
Sept	4.1	2.6	1.4	0	-1.2	-2.6	-4.2
Oct	0.8	0.5	0.3	0	-0.2	-0.5	-0.8
Nov	-2.9	-1.8	-1.0	0	0.8	1.8	3.0
Dec	-5.8	-3.6	-1.9	-0.1	1.7	3.6	6.0

B) NEW RESIDUALS -- old residuals + 1.0x(comparison values)

	L	BR	A	W	B	P	C
Jan	-2.9	1.2	0.7	0.2	0.9	0.9	-1.1
Feb	0.2	3.0	1.2	-0.7	-1.0	-0.4	-2.4
Mar	1.3	1.3	0	-0.5	-1.4	0.3	-1.3
Apr	2.5	0.4	0.7	-0.6	-2.0	-0.9	-0.6
May	1.6	-0.2	0.8	0.4	-1.6	-1.7	0.7
June	0.8	-0.6	1.3	1.2	-0.9	-1.9	-0.1
July	0.2	-2.0	-0.2	1.4	0	-0.3	0.6
Aug	1.0	-1.5	-0.1	0.6	0.5	-0.4	-0.1
Sept	0.7	0	1.0	0.7	-0.2	-1.3	-0.8
Oct	0.1	-0.3	-1.4	-1.1	1.2	0.5	1.2
Nov	-3.1	-1.5	-3.2	-0.6	2.6	2.7	3.1
Dec	-2.9	0.2	-1.2	-0.8	1.5	1.7	1.1

(Cont'd)

exhibit 25 (Cont'd)

C) RESIDUALS AS A BATCH -- unit 0.1°F

to		
2	-3	12
7	-2	94009
21	-1	1043879434513 2
(32)	-0	74056962691214103283 68
41	0	7299234784782665707152
19	1	22336320022571
5	2	567
2	3	01

21 + 22 + 41 = 84, hinges at 21h

Hinges:	-0.9h	and +0.9
Midspread:		1.8h
Sides:	-2.8	and +2.7h
Outside:	4	and 2
(count/20):		(4)

$$\text{clarity} = \frac{87.5-9.8}{1.85} = 42$$ (extreme fits from next panel)

D) SOME VALUES OF THE FIT

Place	Month	Old fit*	New fit**
Laredo	July	93.9	87.5
Laredo	January	53.8	60.5
Caribou	July	56.8	63.4
Caribou	January	16.7	9.8

*As in exhibit 22

**As in this exhibit -- most easily found from:

fit = observed - residual

We have now to go back to the basic definition of the clarity:

$$\frac{\text{highest fit - lowest fit}}{\text{midspread of residuals}}$$

which we now find has increased to 42.

We do have a better fit. Can we make a good picture of it?

18F. Pictures of expanded fits

Let us come back to the general case, where the fit is

$$b + c + d + \frac{cd}{A}$$

where

b = common term (constant)

c = column effect

d = row effect

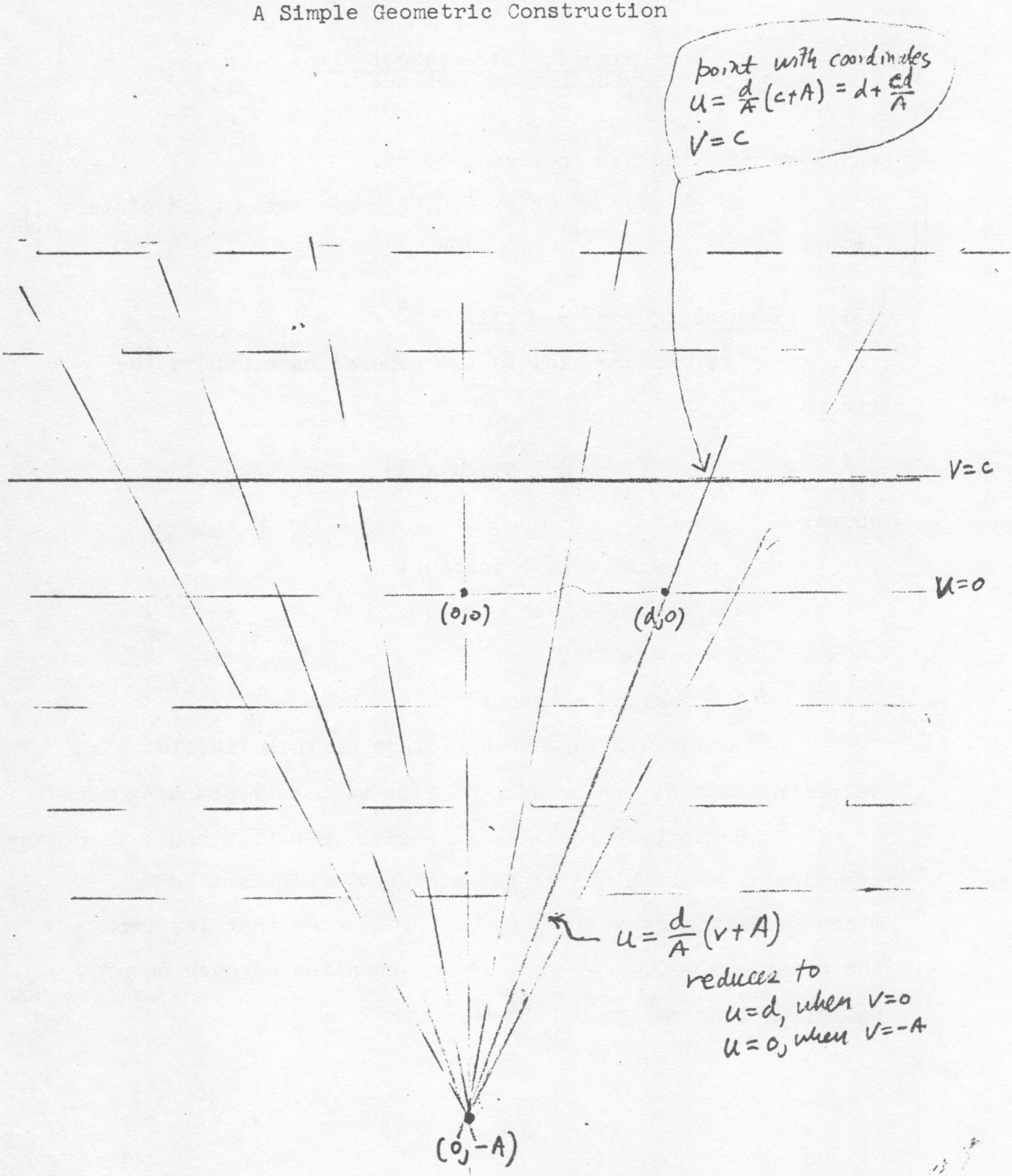
A = scaling constant.

Our job is to make a picture of this labelled with values of tags of c and d, (or with corresponding tags).

Exhibit 26 shows -- in a picture with u and v as ordinary coordinates -- both horizontal lines v = constant and lines radiating from the point $(0,-A)$ -- that is, from the point $u = 0$, $v = -A$. The equation of each such radiating line can be put in the form

exhibit 26 of chapter 18

A Simple Geometric Construction



$$u = \frac{d}{A}(v+A)$$

which is satisfied by $u = 0$, $v = -A$, and which passes through the point $(d,0)$. This line intersects the line $v = c$ in the point with

$$u = \frac{d}{A}(c+A) = d + \frac{cd}{A}$$

This point has, therefore

$$u + v = c + d + \frac{cd}{A}$$

Thus if we replace a vertical line through $(d,0)$ by one through $(d,0)$ twisted to pass through $(o,-A)$, its intersection with $v = c$ will move from

$$u + v = c + d$$

to

$$u + v = c + d + \frac{cd}{A}$$

thus neatly accounting for the additional term in our fit.

We could, of course, interchange u and v if we wished, keeping the vertical lines $u = d$ and passing the other family of lines through $(-A,o)$ and $(0,c)$. Either way, we can easily take care of the added term and still have a straight-line picture. The two fits will be identically the same. In particular, a horizontal line in one will be horizontal in the other.

* back to the East Coast *

Let us apply this to the East Coast monthly temperatures. Exhibit 27 shows the modified construction sheet -- the analog of exhibit 20 of chapter 16. Only solid and dashed lines are needed -- dotted lines are only to remind us of what we had before the additional term was in the fit. Notice that, since A was -56.1, we must have $-A = +56.1$ so that the intersection point comes at a positive month effect.

Clearly, changing from dotted (vertical) lines to solid (radiating) ones has moved one pair of two opposite corners toward lower values of $u + v$, while moving the other pair toward higher values of $u + v$. This is just what we expect such an additional term to do.

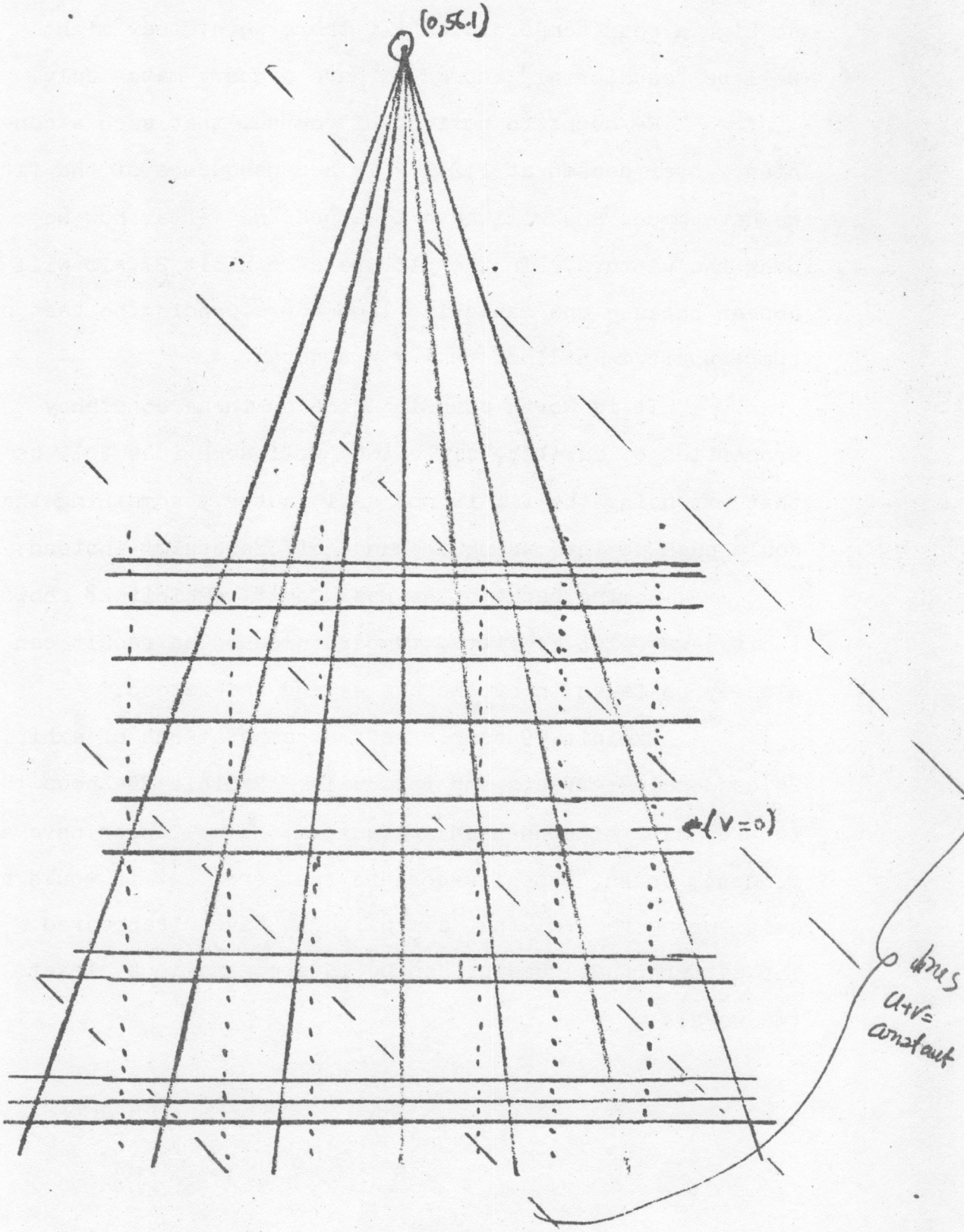
By choosing to have the radiating point at month effect = 56.1, which is nearly as far above July as July is above January, we have avoided any question of obvious paradox. There is no way to get data beyond July.

We might, had we so chosen, have put the radiating point at place effect = 56.1. This would have obvously implied -- were we to be incautious enough to take it seriously -- that once we found an Atlantic Coast place with (annual) place temperature

$$56.1 + 56.1 = 112.2$$

exhibit 27 of chapter 18

Construction figure including additional term
east coast monthly temperatures
(see exhibit 26 and earlier exhibits)



its July temperature would be the same as its January temperature. There seem to be no places with so high a mean temperature. If there were, they might well be "equatorial" enough to have January match July.

We ought to notice, of course, that such a constancy over season at 112.2°F is a consequence of the fit we have made, and will have to appear no matter how we draw the picture. In the picture of exhibit 27, it will appear because the radiating line corresponding to that place temperature is a line $u + v = \text{constant}$.

It is worth spending time on these constancy properties of our fit, since they will sometimes tell us that extending the fit is not satisfactory, something that could push us into using a change of expression instead.

Coming back to the East Coast, exhibit 28 shows the two-way plot of fitted temperatures. The result can clearly be taken in by the eye -- and understood.

Exhibit 29 completes the computations of exhibit 24, condensing effects and residuals. Exhibit 30 shows the two-way plot of condensed residuals. We no longer have any tendency to an "opposite-corners" pattern, but it would be quite wrong to feel that all patterns have disappeared. The pattern that remains can be interpreted in at least two ways:

exhibit 26 of chapter 10

Two-way plot of East Coast temperatures
Extended fit, compare exhibits 2 and 23
(see exhibit 27, and earlier exhibits)

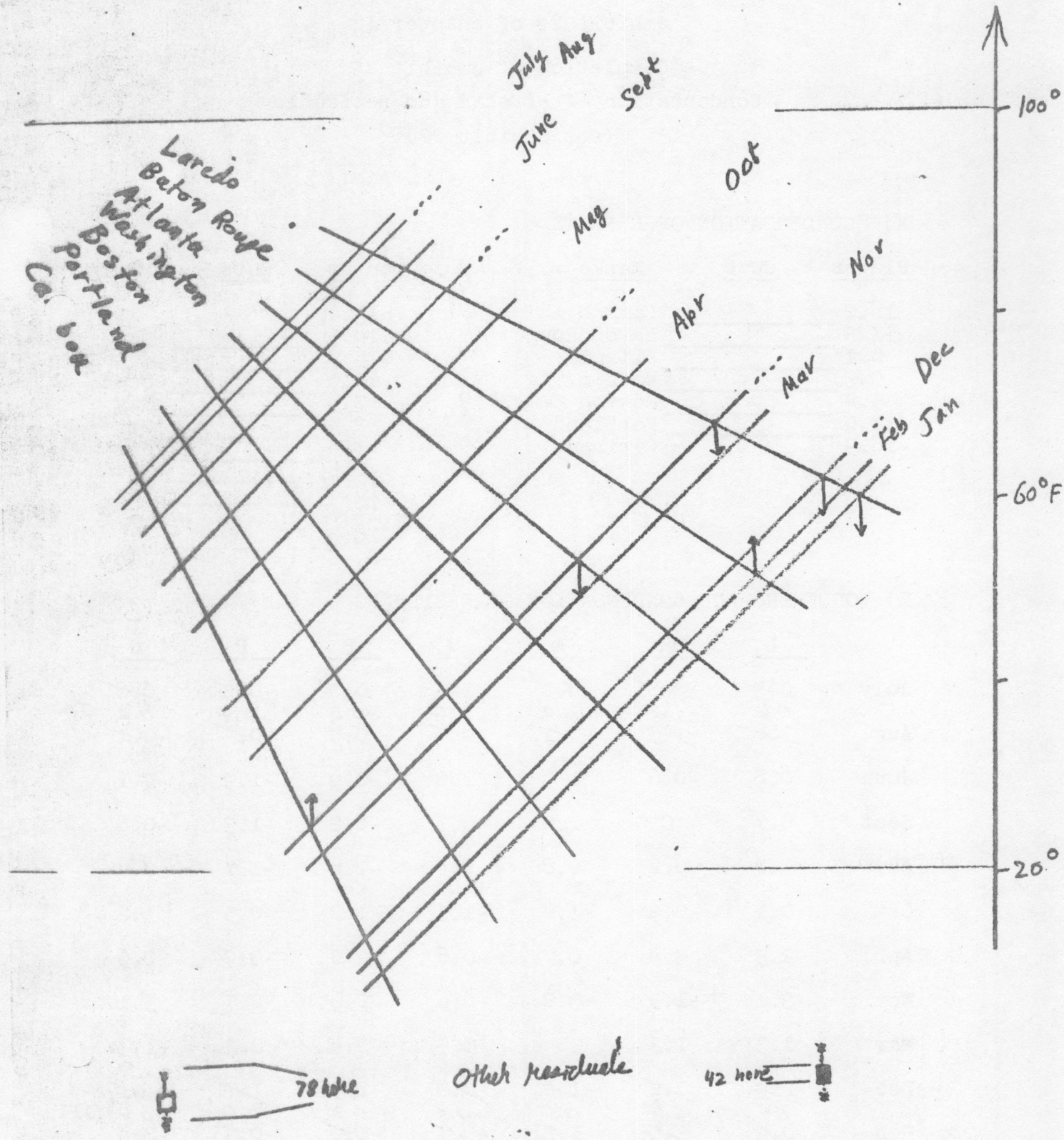


exhibit 29 of chapter 18

Completion of exhibit 25

Condensation of effects and residuals

(see exhibit 22)

A) CONDENSATION OF EFFECTS

Places	Cuts	Marks
18.2		Laredo
11.4		Baton Rouge
6.1		Atlanta
0.3		Wash'n
-5.4		Boston
-11.6		Portland
-18.9		Caribou

Months	Cuts	Marks
19.6		JA
18.5		
15.9		June
12.6		Sept
7.9		May
2.9		Oct
-1.4		Apr
-8.8		Nov
-10.7		Mar
-17.7		DF
-18.7		
-20.5		Jan

B) ORDERING AND CONDENSATION OF RESIDUALS

	L	BR	A	W	B	P	C	
July	~~0.3~~	~~-2.0~~	~~-0.2~~	~~1.4~~	~~0.0~~	~~-0.3~~	~~0.6~~	
	0.6	-1.8	-0.2	1.0	0.2	-0.4	0.2	JA
Aug	~~1.0~~	~~-1.5~~	~~-0.1~~	~~0.6~~	~~0.5~~	~~-0.4~~	~~-0.1~~	
June	0.8	-0.6	1.3	1.2	-0.9	-1.9	-0.1	
Sept	0.7	0	1.0	0.7	-0.2	-1.3	-0.8	
May	1.6	-0.2	0.8	0.4	-1.6	-1.7	0.7	
Oct	0.1	-0.3	-1.4	-1.1	1.2	0.5	1.2	
Apr	2.5	0.4	0.7	-0.6	-2.0	-0.9	-0.6	
Nov	-3.1	-1.5	-3.2	-0.6	2.6	2.7	3.1	
Mar	1.3	1.3	0	-0.5	-1.4	0.3	-1.3	
Dec	~~-2.9~~	~~0.2~~	~~-1.2~~	~~-0.8~~	~~1.5~~	~~1.7~~	~~1.1~~	
	-1.4	1.6	0.0	-0.8	0.2	0.6	-0.6	DF
Feb	~~0.2~~	~~3.0~~	~~1.2~~	~~-0.7~~	~~-1.0~~	~~-0.4~~	~~-2.4~~	
Jan	-2.9	1.2	0.7	0.2	0.9	0.9	-1.1	

(Cont'd)

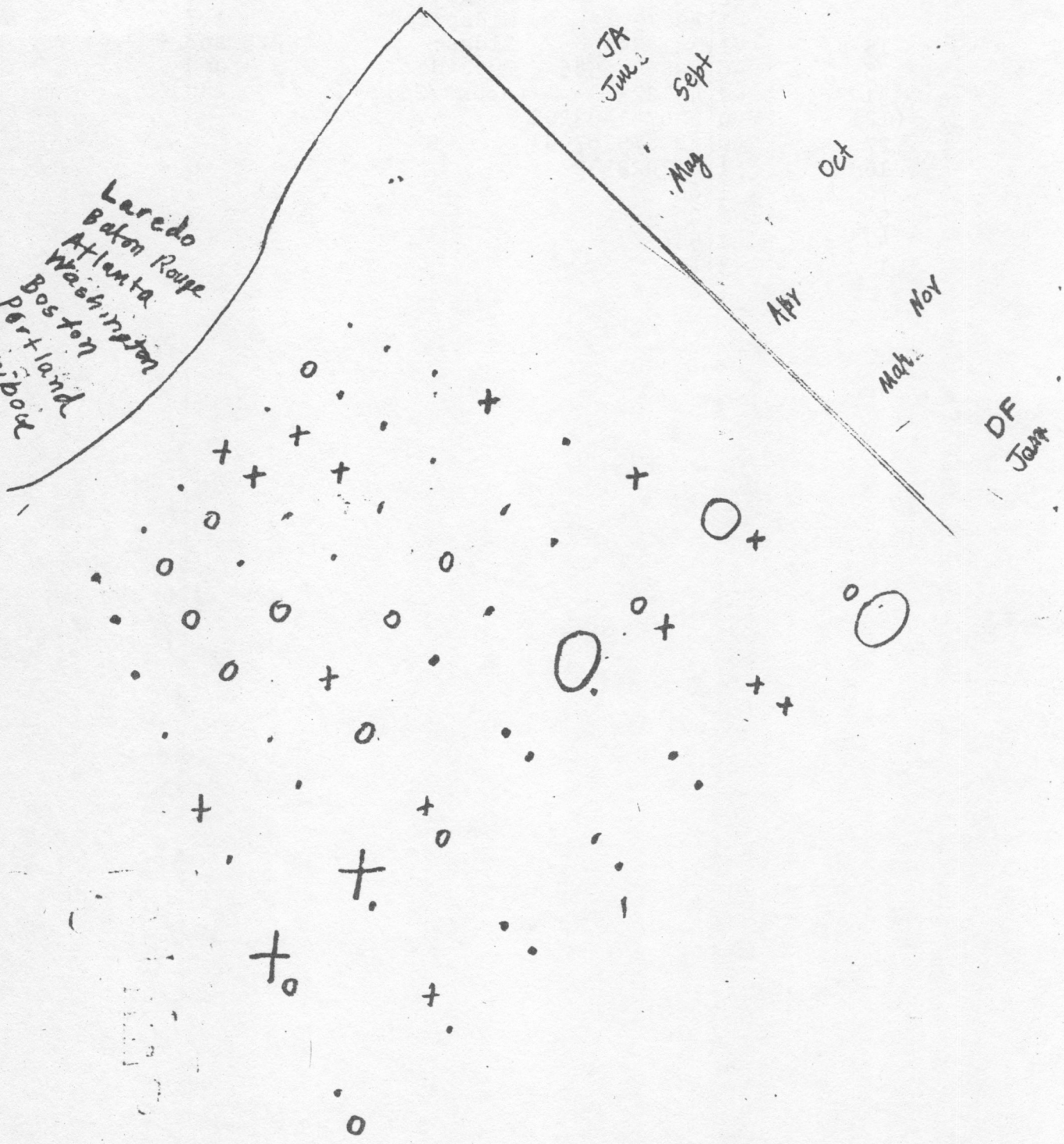
exhibit 29 (Cont'd)

C) CONDENSED RESIDUALS as a BATCH

```
to          (unit = 0.1)
 2          -3|12            31 + 12 + 27 = 70, hinges at 18
 3          -2|9
            -2|              Hinges:          - .8 and + .9
 8          -1|89675         Midspread:            1.7
15          -1|3414341       Sides:           -2.5 and + 2.6
25          -0|6986966586    Outside:          3   and   2
31          -0|421223      (count/20):             (3h)
(12)         0|220741403202
27           0|68787576799
16           1|032022332
 7           1|66
 5           2|0
 4           2|557
 1           3|1
```

exhibit 30 of chapter 18

Two-way Plot of cc Residuals
Extended Fit to East Coast Monthly Temperatures
(see exhibit 28)



- as tendencies to variation along the January-July axis, different places behaving differently.

Thus Baton Rouge, Boston, and Portland appear low in summer and high in winter, while Atlanta and Washington show the reverse trend.

- a tendency to be different, "North of Washington" vs. "South of Washington", different months behaving differently. Thus March, May, June and September are cool in the North and warm in the South; while October, November -- and possibly some mid-winter months -- show the reverse tendency.

Whichever of these -- or whatever other competitor for describing the effect -- is most useful, it would not have been found had we not turned to residuals from a fit at least as careful as this one.

18G. Diagnostic plots need resistant analyses

We have seen that a solid tilt of a residual-vs-comparison value plot means that either a change in expression or an expansion of the fit can be effective. Sometimes, however, an apparently clear and definite tilt can deserve a different interpretation.

* dry/wet ratios *

Exhibit 31 shows data, elementary analysis and comparison values for a 4 by 4 table of ratios of dry to wet weights for wheat cited by cochran. The ratio of 1.035 is clearly impossible, yet as Cochran says "the gross error was at first unnoticed when the data was analyzed arithmetically".

exhibit 31 of chapter 18

Ratio of dry to wet weights for wheat grain

(Dry weight in 1000ths of wet weight)

A) DATA

Block	Nitrogen applied: None	Early	Mid	Late
1	718	732	734	792
2	725	781	725	716
3	704	1 035	763	758
4	726	765	738	761

B) ELEMENTARY ANALYSIS -- by means, essentially as originally done

					eff	fit
1	18	- 78	12	48	- 18	744
2	32	- 22	10	- 21	- 27	737
3	- 67	54	- 30	- 57	53	815
4	18	- 53	08	29	- 10	752
eff	- 44	66	- 22	00	762	00
fit	18	828	740	762	00	- 762

C) COMPARISON VALUES

1	1.0	1.6	.5	.0
2	1.6	2.3	.8	.0
3	3.1	4.6	1.5	.0
4	.6	.9	.3	.0

D) SOURCE

W. G. Cochran 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics 3:22-38, especially Table I at page 27.

Exhibit 32, the corresponding diagnostic plot, shows a strong tilt that is strongly supported by the one point corresponding to the one impossible value. If we cover up this point, however, we still see a strong tilt. So far as the diagnostic plot goes, then, it would seem that we need either to change expression or extend the fit.

The indicated change of expression is to some power like the -31st, which hardly seems reasonable. The extended fit is shown in exhibit 33. The range of the residuals is reduced from .232 to .109 and the midspread from .065 to .043. Clearly the fit has been improved.

Equally clearly -- once we go back and look hard at the original data -- we have improved the fit in the wrong way. We should have been concentrating on the one horrible value, rather than working all over the table.

That we have been thus led astray is not the fault of the diagnostic plot. The question it seeks to answer may be put thus:

> - given the row and column fits, can we do better by including a multiple of the comparison value in the fit -- if so, how much should we include, and to what change in expression might this amount correspond?

That question was answered quite correctly. As exhibit 32 shows so clearly, we can indeed improve the fit considerably. The failure of our procedure lies in the first six words:

given the row and column fits

exhibit 32 of chapter 18

Diagnostic plot for the data of exhibit 31

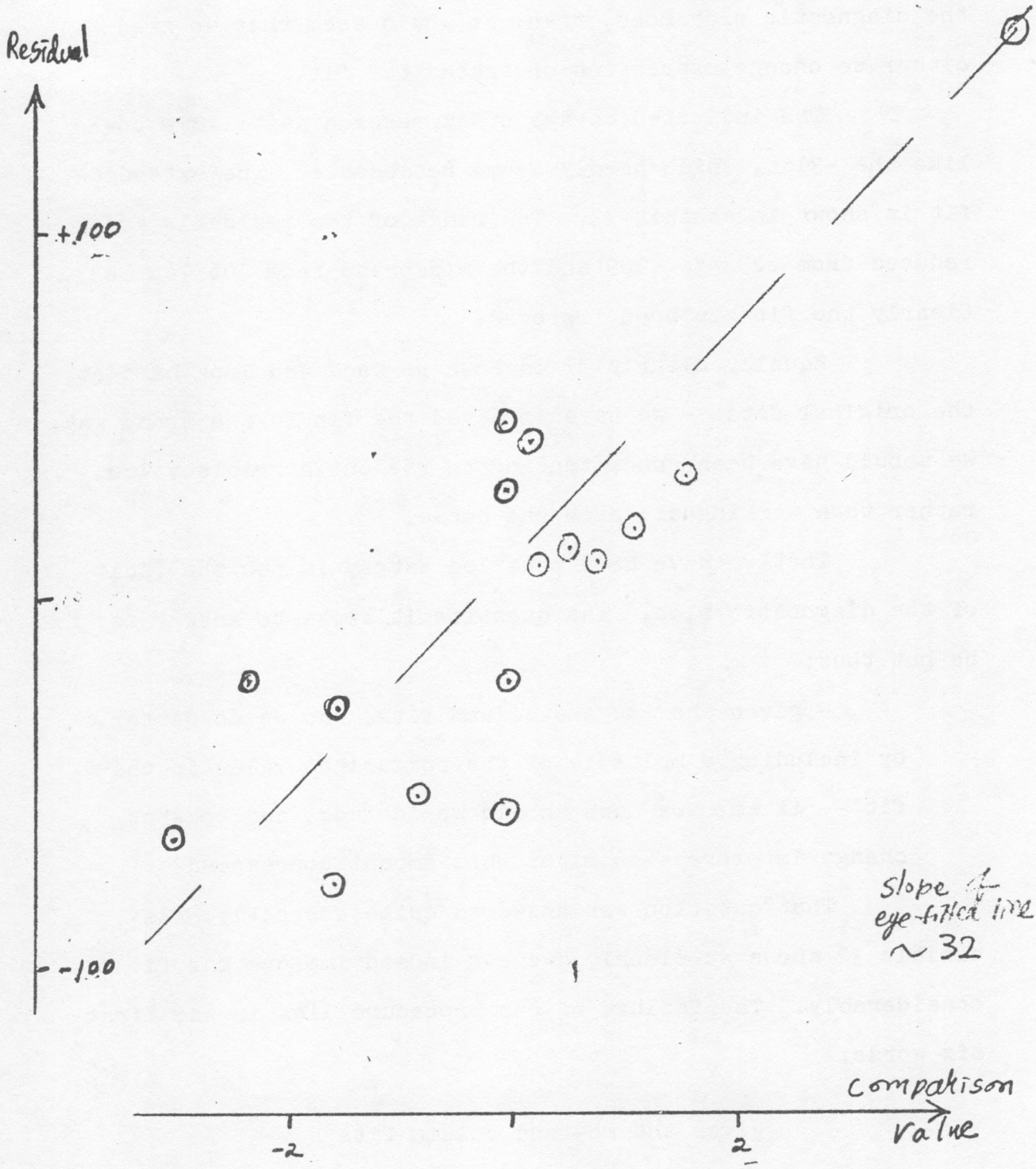


exhibit 33 of chapter 18

Extending the fit of exhibit 31

A) 32 times the COMPARISON VALUES

	None	Early	Middle	Late
1	32	-51	16	0
2	51	- 74	26	0
3	-100	148	- 48	0
4	19	- 29	10	0

B) NEW RESIDUALS -- panel B of exhibit 3 MINUS A above

1	- 14	- 27	- 4	48
2	- 19	52	- 16	- 21
3	33	06	18	- 57
4	- 01	- 24	- 2	+ 29

C) RESIDUALS as BATCHES -- hinges at depth 4h

B of exhibit 3

H	H
.05	
.04	8
.03	2
.02	9
.01	8820
.00	8
-.00	
-.01	
-.02	21
-.03	0
-.04	
-.05	37
-.06	7
-.07	8

midspread = .065

Hinges: +23h and - 41h
Sides: +88h and -106h
Outside: 1 and 0
(count/20): (h)

B above

H	H
.05	2
.04	8
.03	3
.02	9
.01	8
.00	6
-.00	142
-.01	496
-.02	741
-.03	
-.04	
-.05	7
-.06	
-.07	

midspread = .043

Hinges: +23h and -20
Sides: +66h and -63
Outside: 0 and 0
(count/20): (h)

We were in trouble once these fits were fixed. So far we have done nothing to get out of this trouble. The difficulty is in the lack of resistance in the process producing the initial fits.

Exhibit 34 shows the results of analyzing the original data in the way chosen in chapter 19 -- stripe medians separately and then a look at pomedians, which turn out to be zero. The comparison values are now even smaller, the second largest size being just over .2. (Note that the fitted column effects have been roughly halved, while the row effects have been roughly divided by 6. Thus comparison values ought only be about one-twelfth of their original size.) This smallness ought to warn us that we cannot afford to do much by direct re-expression -- a few times .2 can only amount to 1 or 2, while about half our residuals are of size 10 or greater.

Suppose, however, that we miss this cue, and go on to the diagnostic plot. Exhibit 35 shows the result -- one point far away and the remaining 15 behaving in a rather non-descript way. The message -- one bad point, no clear desire for re-expression or fit-extension -- is now quite clear.

We now have this example free of misinterpretation. Do we have it free of unneeded distortion? Probably not. If we look at the row and column of residuals containing the wild value, we see an excess of minuses among the residuals and

exhibit 34 of chapter 18 -- dry/wet ratios in 1000ths

Median analysis of the wheat drying data
(see exhibit 31)

A) TWO-WAY ANALYSIS by MEDIANS -- by columns and then by rows; stripe pomedians now zero

	Nitrogen applied					
Block	None	Early	Mid	Late	eff	fit
1	-1	-38	1	25	-3	750
2	5	10	-9	-52	-2	751
3	-26	254	19	-20	8	761
4	1	-11	-1	8	3	756
eff	-31	20	-17	17	753	
fit	722	773	736	770		

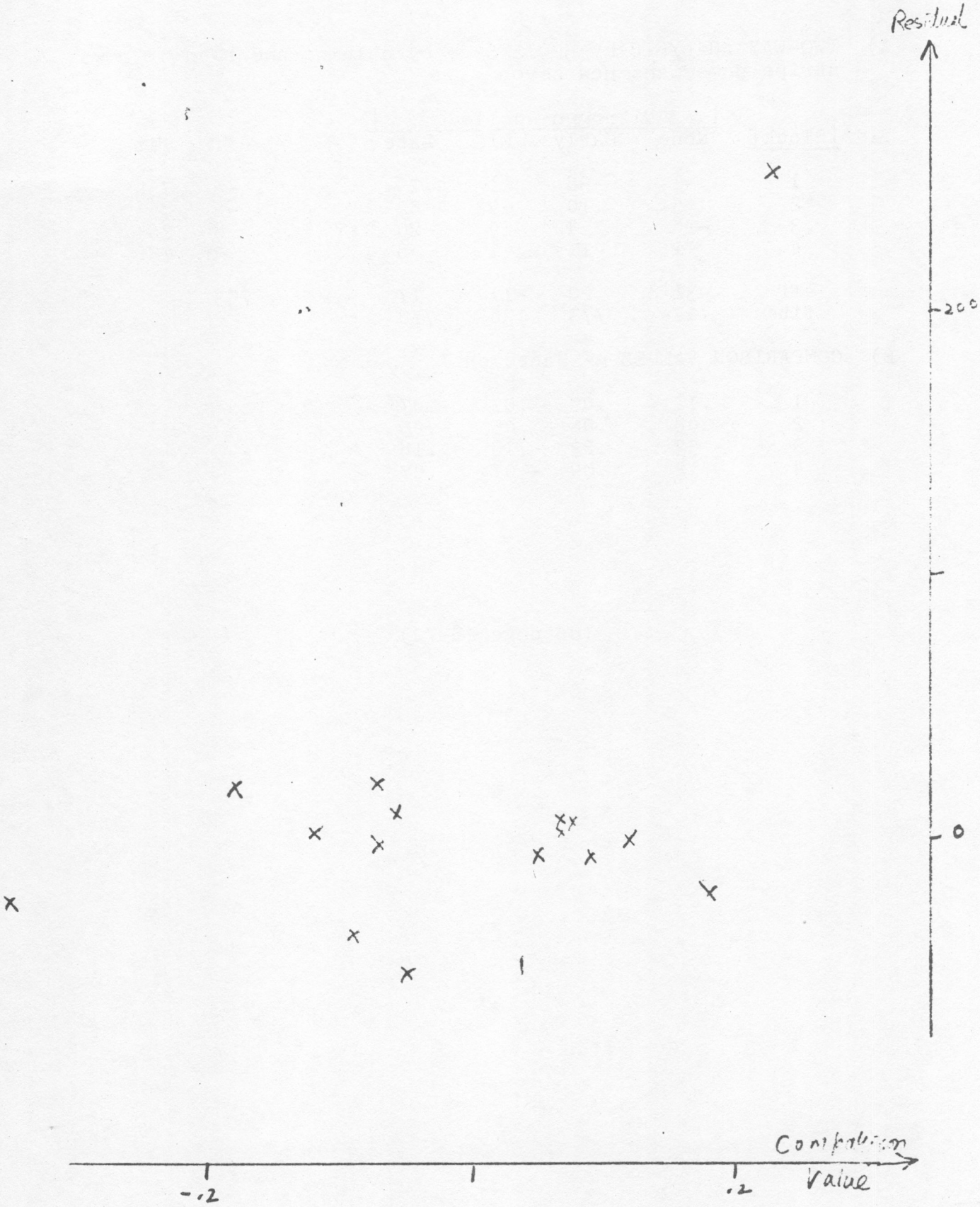
B) COMPARISON VALUES -- based on A

1	.12	-.09	.07	-.07
2	.08	-.06	.05	-.05
3	-.33	.23	-.18	.18
4	-.12	.09	-.07	.07

(no page 18-72)

exhibit 35 of chapter 18 -- dry/wet ratios in 1000ths

Diagnostic plot based on median analysis
(from panels A and B of exhibit 34)

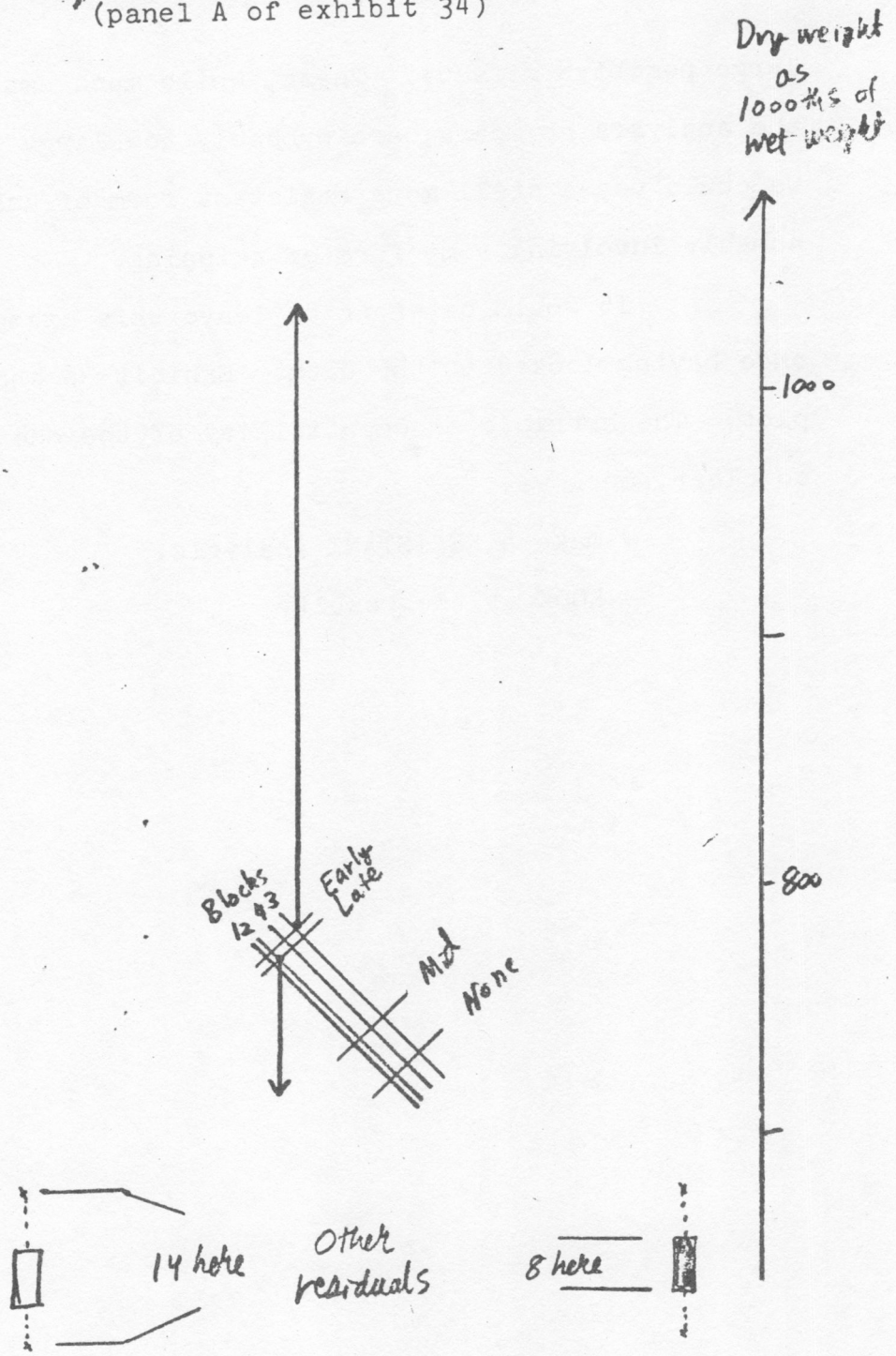


large positive effects. These, while much smaller than for the analysis by means, are probably too large for comfort. We could use a still more resistant form of analysis, presumably involving some form of skipping.

It would be wrong to leave this example without once having looked at the data. Exhibit 36 shows the two-way plot. The horrible incompatibility of the one residual is so, so clear, once we:

- make a RESISTANT analysis.
- LOOK at the result.

Two-way plot of the resistant analysis
(panel A of exhibit 34)



18H. How far have we come?

This chapter has taken us down two roads:

- obtaining better fits by adopting a better expression of the response

- obtaining better fits by adding certain "natural" terms. (In chapter 17, we already had improved a fit by adding special terms. The difference here is that these "natural" terms are available for every two-way table since their parts are in every two-way analysis.

These roads proved more closely related than they might have seemed. The tie between them is the quantity

$$\text{comparison value} = \frac{(\text{row effect})(\text{column effect})}{(\text{common term})}$$

On the one hand, the tilt of residuals, when plotted against comparison values, suggests how much of change in expression to try. (Near 0.5, try root. Near 1.0, try log. Near 2.0 try reciprocal. Near -1, try square. Near -2, try cube.) On the other, the natural first step in extending the fit is to include the corresponding multiple of the comparison value in the fit.

We found it convenient, especially when making two-way plots, to write this additional term as

$$\frac{(\text{row effect})(\text{column effect})}{A}$$

where A is in the same units as all responses and effects.

To make a two-way plot of the extended fit, we have only to replace one set of parallel lines by a fan of lines coming out from a point. Two choices are alternative, equivalent, and simple:

- retain the lines v = constant, and draw the fan through u = 0, v = -A and the intersections of the old lines u = constant with v = 0.

- retain the lines u = constant, and draw the fan through u = -A, v = 0 and the intersections of the old lines v = constant with u = 0.

A single wild value, in a two-way table analyzed by means, can -- when we are incautious and fit by means -- give a clear appearance of tilt in a diagnostic plot of residuals vs. comparison values. This is due to lack of resistance in fitting and can be overcome by using a more resistant fitting process, as by medians and pomedians. (The usefulness of still more resistance is hinted at by our example.)

*****

18K. <u>An interpretation</u> (background)

The appearance of

$$u \cdot v$$

where

$$u = \text{row effect}$$
$$v = \text{column effect}$$

as the extended fit may seem quite new and unusual. There is a way to look at matters, however, in which it is "just more of the same".

When fitting a straight line

$$a + bx$$

to y, we felt the natural next step was to start squaring the fit, leading us to -- as an additional term -- any multiple of

$$(a+bx)^2 = b^2 \cdot x^2 + 2ab \cdot x + a^2$$

We then found that modifying this by adding or subtracting anything of the form already fitted, thus reaching

$$\pm\, b^2 \cdot (x^2 - dx - e)$$

for any d and e,was an equivalent approach.

In a two-way table we began by fitting

$$u + v + c$$

whose square is

$$(u+v+c)^2 = 2uv + u^2 + 2c\cdot u + v^2 + 2c\cdot v + c^2$$

The last five terms, once bracketed like this

$$(u^2+2c\cdot u) + (v^2+2c\cdot v) + c^2$$

are easily seen to be of the form

row term + col term + constant

namely what is already fitted. Accordingly we ought to expect that using

$$\pm\ 2\cdot uv$$

or any other multiple of uv as an additional fit ought to give the same results as using

$$\pm\ (u+v+c)^2$$

Thus comparison values can be considered just another case of squaring the previous fit and seeing what adding it will do.